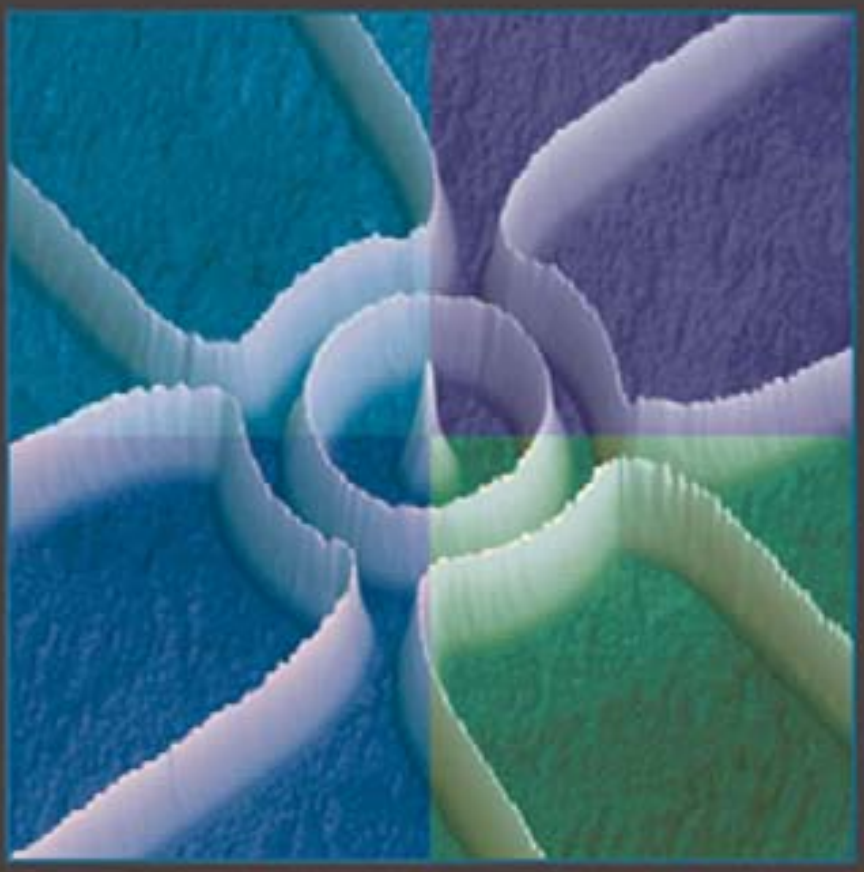
OXFORD

Semiconductor Nanostructures

Quantum States and Electronic Transport



Thomas Ihn

SEMICONDUCTOR NANOSTRUCTURES

This page intentionally left blank

Semiconductor Nanostructures

Quantum States and Electronic Transport

Thomas Ihn
Solid State Physics Laboratory, ETH Zurich

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi New Delhi
Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal
Singapore South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2010

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Printed in the UK
on acid-free paper by
by CPI Antony Rowe, Chippenham, Wiltshire

ISBN 978–0–19–953442–5 (Hbk.)
ISBN 978–0–19–953443–2 (Pbk.)

1 3 5 7 9 10 8 6 4 2

Preface

This book is based on the lecture notes for the courses *Semiconductor Nanostructures* and *Electronic Transport in Nanostructures* that the author gives regularly at the physics department of ETH Zurich. The course is aimed at students in the fourth year who have already attended the introductory lectures Physics I–IV, theoretical lectures in electrodynamics, classical and quantum mechanics, and a course Introduction to Solid State Physics. The course is also attended by PhD students within their PhD programme, or by others working in the field of semiconductor nanostructures or related scientific areas. Beyond the use of the material contained in this book as the basis for lectures, it has become a popular reference for researchers in a number of research groups at ETH working on related topics. This book is therefore primarily intended to be a textbook for graduate students, PhD students and postdocs specializing in this direction.

In order to acquire the knowledge about semiconductor nanostructures needed to understand current research, it is necessary to look at a considerable number of aspects and subtopics. For example, we have to answer questions like: which semiconducting materials are suitable for creating nanostructures, which ones are actually used, and which properties do these materials have? In addition, we have to look at nanostructure processing techniques: how can nanostructures actually be fabricated? A further topic is the historical development of this modern research field. We will have to find out how our topic is embedded in the physical sciences and where we can find links to other branches of physics. However, at the heart of the book will be the physical effects that occur in semiconductor nanostructures in general, and more particularly on electronic transport phenomena.

Using this book as the basis for a course requires selection. It would be impossible to cover all the presented topics in depth within the fourteen weeks of a single semester given two hours per week. The author regards the quantization of conductance, the Aharonov–Bohm effect, quantum tunneling, the Coulomb blockade, and the quantum Hall effect as the five fundamental transport phenomena of mesoscopic physics that need to be covered. As a preparation, Drude transport theory and the Landauer–Büttiker description of transport are essential fundamental concepts. All this is based on some general knowledge of semiconductor physics, including material aspects, fabrication, and elements of band structure. This selection, leaving out a number of more specialized and involved

topics would be a solid foundation for a course aimed at fourth year students.

The author has attempted to guide the reader to the forefront of current scientific research and also to address some open scientific questions. The choice and emphasis of certain topics do certainly follow the preference and scientific interest of the author and, as illustrations, his own measurements were in some places given preference over those of other research groups. Nevertheless the author has tried to keep the discussions reasonably objective and to compile a basic survey that should help the reader to seriously enter this field by doing his or her own experimental work.

The author wishes to encourage the reader to use other sources of information and understanding along with this book. Solving the exercises that are embedded in the chapters and discussing the solutions with others is certainly helpful to deepen understanding. Research articles, some of which are referenced in the text, or books by other authors may be consulted to gain further insight. You can use reference books, standard textbooks, and the internet for additional information. Why don't you just start and type the term 'semiconductor nanostructures' into your favorite search engine!

Thomas Ihn,
Zurich, January 2009

Acknowledgements

I want to thank all the people who made their contribution to this book, in one way or other. I thank my family for giving me the freedom to work on this book, for their understanding and support. I thank all the colleagues who contributed with their research to the material presented. I thank my colleagues at ETH who encouraged me to tackle this project. Many thanks go to all the students who stimulated the contents of the book by their questions and comments, who found numerous mistakes, and who convinced me that it was worth the effort by using my previous lecture notes intensively.

I wish to acknowledge in particular those present and former colleagues at ETH Zurich who contributed unpublished data, drawings, or other material for this book:

Andreas Baumgartner, Christophe Charpentier, Christoph Ellenberger, Klaus Ensslin, Andreas Fuhrer, Urszula Gasser, Boris Grbič, Johannes Güttinger, Simon Gustavsson, Renaud Leturcq, Stephan Lindemann, Johannes Majer, Francoise Molitor, Hansjakob Rusterholz, Jörg Rychen, Roland Schleser, Silke Schön, Volkmar Senz, Ivan Shorubalko, Martin Sigrist, Christoph Stampfer, Tobias Vančura.

This page intentionally left blank

Contents

This page intentionally left blank

Introduction

1

1.1 A short survey

Nanostructures in physics. How is the field of semiconductor nanostructures embedded within more general topics which the reader may already know from his or her general physics education? Figure 1.1 is a graphical representation that may help. Most readers will have attended a course in solid state physics covering its basics and some of its important branches, such as magnetism, superconductivity, the physics of organic materials, or metal physics. For this book, the relevant branch of solid sate physics is semiconductor physics. Particular aspects of this branch are materials, electrical transport properties of semiconductors and their optical properties. Other aspects include modern semiconductor devices, such as diodes, transistors and field-effect transistors.

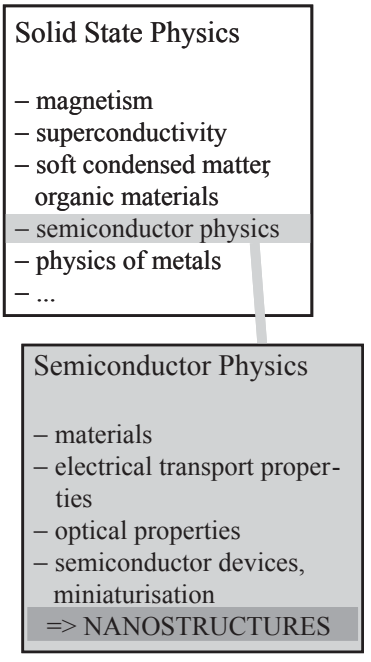


Fig. 1.1 Schematic representation showing how the field of semiconductor nanostructures has emerged as a special topic of solid state physics.

Miniaturization of electronic devices in industry and research. We all use modern electronics every day, sometimes without being aware of it. It has changed life on our planet during the past fifty years enormously. It has formed an industry with remarkable economical success and a tremendous influence on the world economy. We all take the availability of computers with year by year increasing computing power for granted. The reason for this increase in computer power is, among other things, the miniaturization of the electronic components allowing us to place a steadily increasing amount of functionality within the same area of a computer chip. The decreasing size of transistors also leads to decreasing switching times and higher clock frequencies. Today's silicon-based computer processors host millions of transistors. The smallest transistors, fabricated nowadays in industrial research laboratories have a gate length of only 10 nm.

Of course, this trend towards miniaturization of devices has also affected semiconductor research at universities and research institutes all over the world and has inspired physicists to perform novel experiments. On one hand they benefit from the industrial technological developments which have established materials of unprecedented quality and innovative processing techniques that can also be used in modern research. On the other hand, physicists are interested in investigating and understanding the physical limits of scalability towards smaller and smaller devices, and, eventually, to think about novel device concepts beyond the established ones. Can we realize a transistor that switches with single electrons? Are the essentially classical physical concepts that govern

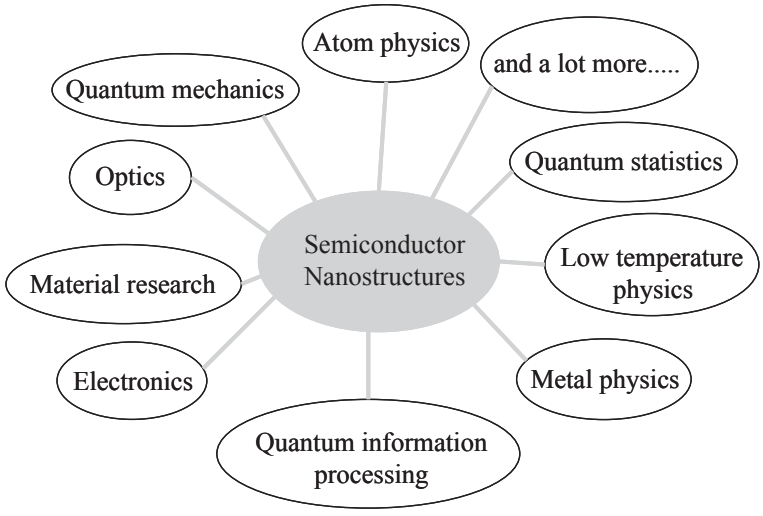


Fig. 1.2 The physics of semiconductor nanostructures is related to many other areas of physics.

the operation of current transistors still applicable for such novel devices? Do we have to take quantum effects into account in such small structures? Can we develop new operating principles for semiconductor devices utilizing quantum effects? Can we use the spin of the electrons as the basis for spintronic devices?

All these highly interesting questions have been the focus of research in industry, research institutes, and universities for many years. In the course of these endeavors the field of semiconductor nanostructures was born around the mid 1980s. Experiments in this field utilize the technological achievements and the quality of materials in the field of semiconductors for fabricating structures which are not necessarily smaller than current transistors but which are designed and investigated under conditions that allow quantum effects to dominate their properties. Necessary experimental conditions are low temperatures, down to the millikelvin regime, and magnetic fields up to a few tens of tesla. A number of fundamental phenomena has been found, such as the quantization of conductance, the quantum Hall effect, the Aharonov–Bohm effect and the Coulomb-blockade effect. In contrast, quantum phenomena play only a minor role in today's commercial semiconductor devices.

Nanostructure research and other branches of physics. The physics of semiconductor nanostructures has a lot in common with other areas of physics. Figure 1.2 is an attempt to illustrate some of these links. The relations with materials science and electronics have already been mentioned above. Beyond that, modern semiconductor electronics is an integrated part of measurement equipment that is being used for the measurement of the physical phenomena. The physics of low temperatures is very important for experimental apparatus such as cryostats which are necessary to reveal quantum phenomena in semiconductor nanostructures. Quantum mechanics, electrodynamics and quantum

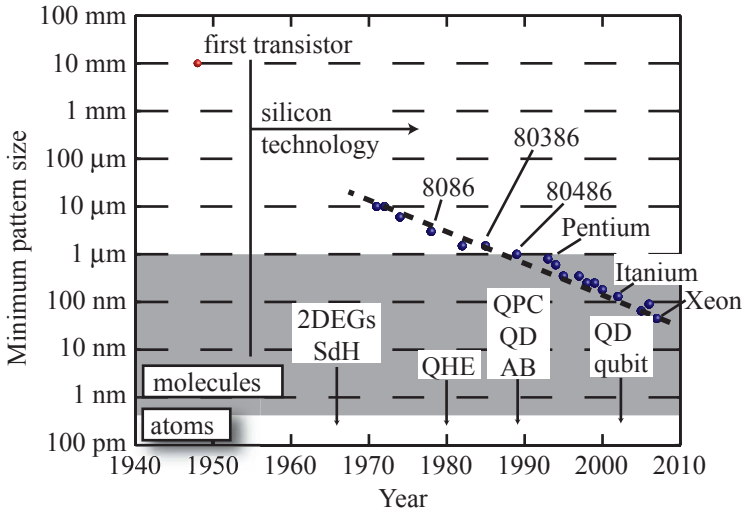


Fig. 1.3 Development of the minimum pattern sizes in computer processor chips over time. Data on Intel processors were compiled from Intel publications. The dashed line represents the prediction of Moore's law, i.e., an exponential decrease of pattern size over time. Abbreviations in the bottom part of the chart indicate milestones in semiconductor nanostructure research, namely, 2DEGs: two-dimensional electron gases, SdH: Shubnikov–de Haas effect, QHE: quantum Hall effect, QPC: quantum point contact (showing conductance quantization), QD: quantum dot (Coulomb blockade), AB: Aharonov–Bohm effect, QD qubit: quantum dot qubit. This shows the close correlation between industrial developments and progress in research.

statistics together form the theoretical basis for the description of the observed effects. From metal physics we have inherited models for diffusive electron transport such as the Drude model of electrical conduction. Analogies with optics can be found, for example, in the description of conductance quantization in which nanostructures act like waveguides for electrons. We use the terms 'modes', 'transmission', and 'reflection' which are also used in optics. Some experiments truly involve electron optics. The field of zero-dimensional structures, also called quantum dots or artificial atoms, has strong overlap with atom physics. The fact that transistors are used for classical information processing and the novel opportunities that nanostructures offer have inspired researchers to think about new quantum mechanical concepts for information processing. As a result, there is currently a very fruitful competition between different areas of physics for the realization of certain functional units such as quantum bits (called qubits) and systems of qubits. The field of semiconductor nanostructures participates intensely in this competition. Reading this book you will certainly find many other relations with your own previous knowledge and with other areas of physics.

History and Moore's law. Historically, the invention of the transistor by Shockley, Bardeen, and Brattain, at that time at the Bell laboratories, was a milestone for the further development of the technological use of semiconductors. The first pnp transistor was developed in 1949 by Shockley. In principle it already worked like today's bipolar transistors. Since then miniaturization of semiconductor devices has made enormous progress. The first transistors with a size of several millimeters had already been scaled down by 1970 to structure sizes of about 10 µm. Since then, miniaturization has progressed exponentially as predicted by Moore's law (see Fig. 1.3). With decreasing structure size the number

of electrons participating in transistor switching decreases accordingly. If Moore's law continues to be valid, industry will reach structure sizes of the order of the electron's wavelength within the next decade. There is no doubt that the importance of quantum effects will tend to increase in such devices.

The size of semiconductor nanostructures. The world of nanostructures starts below a characteristic length of about 1 µm and ends at about 1 nm. Of course, these limits are not strict and not always will all dimensions of a nanostructure be within this interval. For example, a ring with a diameter of 5 µm and a thickness of 300 nm would certainly still be called a nanostructure. The word 'nano' is Greek and means 'dwarf'. Nanostructures are therefore 'dwarf-structures'. They are frequently also called *mesoscopic systems*. The word 'meso' is again Greek and means 'in between', 'in the middle'. This expresses the idea that these structures are situated between the macroscopic and the microscopic world. The special property of structures within this size range is that typically a few length scales important for the physics of these systems are of comparable magnitude. In semiconductor nanostructures this could, for example, be the mean free path for electrons, the structure size, and the phase-coherence length of the electrons.

Beyond the nanostructures lies the atomic world, starting with macromolecules with a size below a few nanometers. *Carbon nanotubes*, small tubes of a few nanometers in diameter made of graphene sheets, are at the boundary between nanostructures and macromolecules. They can reach lengths of a few micrometers. Certain types of these tubes are metallic, others semiconducting. Their interesting properties have made them very popular in nanostructure research of recent years.

Electronic transport in nanostructures. The main focus of this book is the physics of electron transport in semiconductor nanostructures including the arising fundamental quantum mechanical effects. Figure 1.4 shows a few important examples belonging to this theme. Measuring the electrical resistance, for example, using the four-terminal measurement depicted schematically at the top left is the basic experimental method. The quantum Hall effect (bottom left) is a phenomenon that arises in two-dimensional electron gases. It is related to the conductance quantization in a quantum point contact. Another effect that arises in diffusive three-, two-, and one-dimensional electron gases is the so-called weak localization effect (top middle). Its physical origin can be found in the phase-coherent backscattering of electron waves in a spatially fluctuating potential. This effect is related to the Aharonov–Bohm effect in ring-like nanostructures (top right). A characteristic effect in zero-dimensional structures, the quantum dots, is the Coulomb-blockade effect. Its characteristic feature is the sharp resonances in the conductance as the gate voltage is continuously varied. These resonances are related to the discrete energy levels and to the quantization of charge in this many-electron droplet. While the summary of effects shown in Fig.

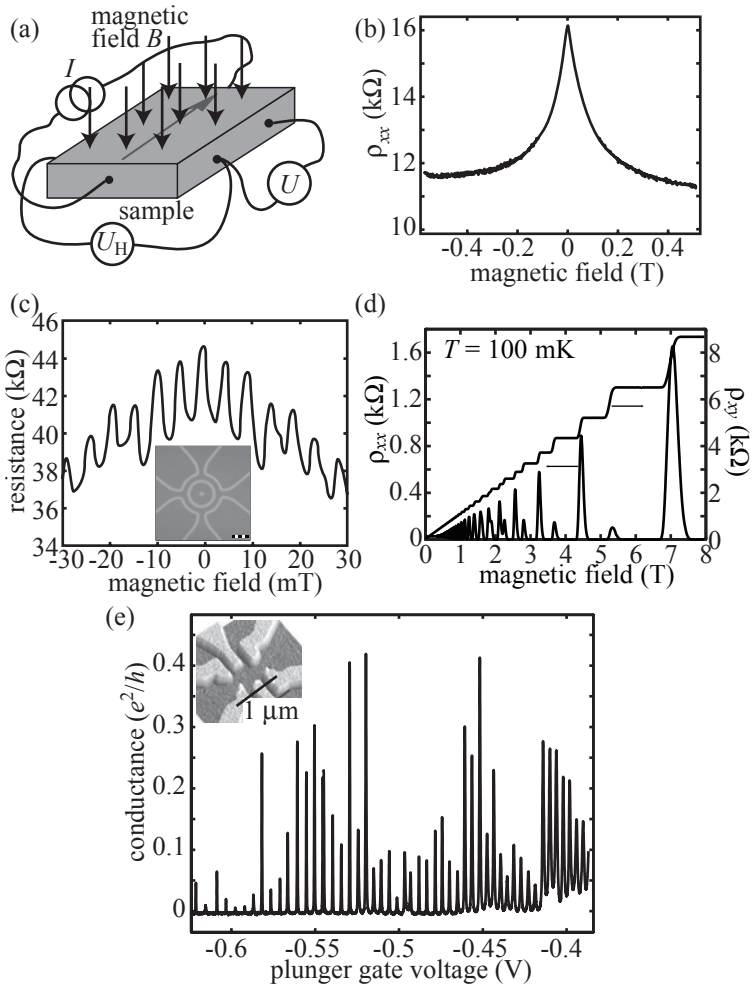


Fig. 1.4 Summary of important quantum transport phenomena in semiconductor nanostructures. (a) Schematic drawing of a four-terminal resistance measurement. (b) Weak localization effect in a diffusive two-dimensional electron gas, which is related to the AharonovBohm effect shown in (c). (c) Aharonov-Bohm effect in a quantum ring structure. (d) The longitudinal- and the Hall-resistivity of a two-dimensional electron gas in the quantum Hall regime. (e) Conductance of a quantum dot structure in the Coulomb-blockade regime.

1.4 cannot be complete, it shows the rich variety of transport phenomena which makes the field of semiconductor nanostructures particularly attractive.

1.2 What is a semiconductor?

The term 'semiconductor' denotes a certain class of solid materials. It suggests that the electrical conductivity is a criterion for deciding whether a certain material belongs to this class. We will see, however, that quantum theory provides us with an adequate description of the band structure of solids and thereby gives a more robust criterion for the distinction between semiconductors and other material classes.

Resistivity and conductivity. The electrical conductivity of solid materials varies over many orders of magnitude. A simple measure-

ment quantity for the determination of the conductivity is the electrical resistance R which will be more thoroughly introduced in section 10.1 of this book. If we consider a block of material with length L and cross-sectional area A, we expect the resistance to depend on the actual values, i.e., on the geometry. By defining the (specific) resistivity

$$\rho = R\frac{A}{L}$$

we obtain a geometry-independent quantity which takes on the same value for samples of different geometries made from the same material. The resistivity is therefore a suitable quantity for the electrical characterization of the material. The (specific) conductivity σ is the inverse of the resistivity, i.e.,

$$\sigma = \rho^{-1}.$$

Empirically we can say that metals have large conductivities, and insulators small, while semiconductors are somewhere in between. Typical numbers are shown in Table 1.1.

Table 1.1 Typical resistivities of materials at room temperature.

Material	ρ (Ωcm)
Insulators	$\sim 10^{14}$
Macor (ceramic)	
SiO_2 (quartz)	
Al_2O_3 (sapphire)	
Semiconductors	$10^{-2} - 10^{9}$
Metals	$\sim 2 \times 10^{-6}$
Cu	1.7×10^{-6}
Al	2.6×10^{-6}
Au	2.2×10^{-6}

Temperature dependence of the resistance. The temperature dependence of the electrical resistance is a good method for distinguishing metals, semiconductors and insulators.

The specific resistivity of metals depends weakly and linearly on temperature. When a metal is cooled down from room temperature, electron–phonon scattering, i.e., the interaction of electrons with lattice vibrations, loses importance and the resistance goes down [see Fig. 1.5(a)]. At very low temperatures T, the so-called Bloch–Grüneisen regime is reached, where the resistivity shows a T^5-dependence and goes to a constant value for $T \to 0$. This value is determined by the purity of, and number of defects in, the involved material. In some metals this 'standard' low-temperature behavior is strongly changed, for example, by the appearance of superconductivity, or by Kondo-scattering (where magnetic impurities are present).

In contrast, semiconductors and insulators show an exponential dependence of resistivity on temperature. The resistance of a pure high-quality semiconductor increases with decreasing temperature and diverges for $T \to 0$ [cf., Fig. 1.5(b)]. The exact behavior of the temperature dependence of resistivity depends, as in metals, on the purity and on the number of lattice defects.

Band structure and optical properties. A very fundamental property that semiconductors share with insulators is their band structure. In both classes of materials, the valence band is (at zero temperature) completely filled with electrons whereas the conduction band is completely empty. A band gap E_g separates the conduction band from the valence band [see Fig. 1.6(a)]. The Fermi level E_F is in the middle of the band gap.

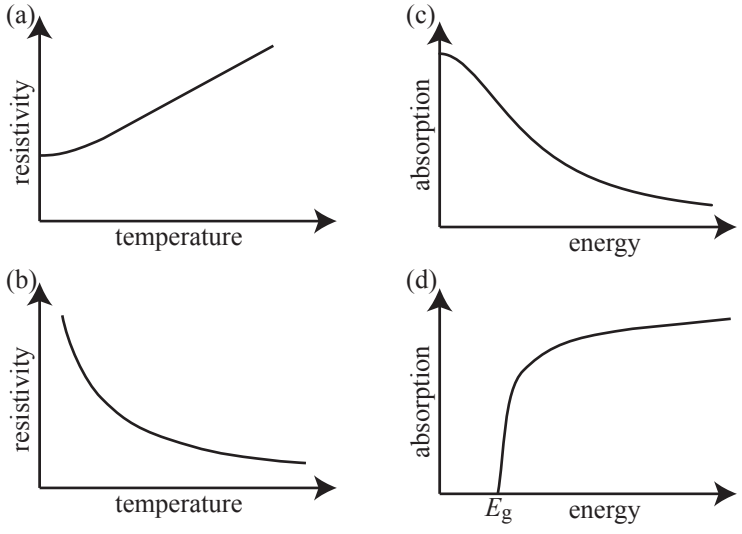


Fig. 1.5 Left: Characteristic temperature dependence of the resistivity (a) of a metal, (b) of a semiconductor. Right: Characteristic optical absorption as a function of photon energy (c) of a metal, (d) of a semiconductor.

This property distinguishes semiconductors and insulators from metals, in which a band gap may exist, but the conduction band is partially filled with electrons up to the Fermi energy E_F and the lowest electronic excitations have an arbitrarily small energy cost [Fig. 1.6(b)].

The presence of a band gap in a material can be probed by optical transmission, absorption, or reflection measurements. Roughly speaking, semiconductors are transparent for light of energy below the band gap, and there is very little absorption. As depicted in Fig. 1.5(d), at the energy of the band gap there is an absorption edge beyond which the absorption increases dramatically. In contrast, metals show a finite absorption at arbitrarily small energies due to the free electrons in the conduction band [Fig. 1.5(c)].

Semiconductors can be distinguished from insulators only by the size of their band gap. Typical gaps in semiconductors are between zero and 3 eV. However, this range should not be seen as a strict definition of semiconductors, because, depending on the context, even materials with larger band gaps are often called semiconductors in the literature. The band gaps of a selection of semiconductors are tabulated in Table 1.2.

Table 1.2 Band gaps (in eV) of selected semiconductors.

Si	Ge	GaAs	AlAs	InAs
1.1	0.7	1.5	2.2	0.4

Doping of semiconductors. A key reason why semiconductors are technologically so important is the possibility of changing their electronic properties enormously by incorporating very small amounts of certain atoms that differ in the number of valence electrons from those found in the pure crystal. This process is called doping. It can, for example, lead to an extreme enhancement of the conductivity. Tailored doping profiles in semiconductors lead to the particular properties utilized in semiconductor diodes for rectifying currents, or in bipolar transistors for amplifying and switching.

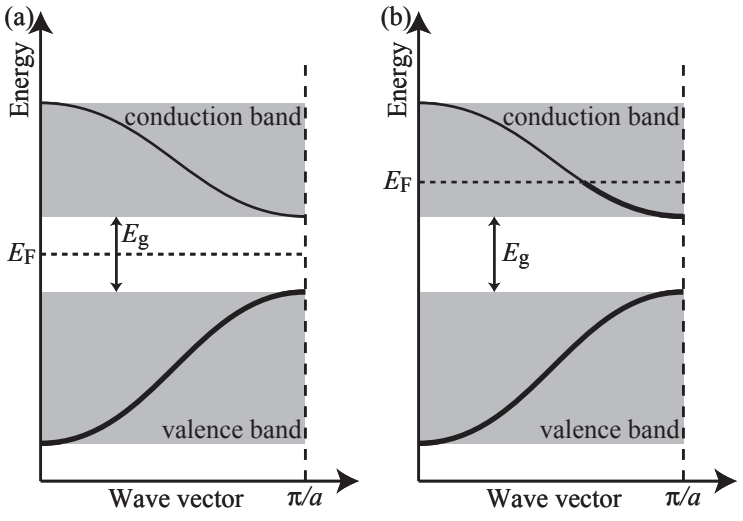


Fig. 1.6 Schematic representation of band structure within the first Brillouin zone, i.e., up to wave vector π/a, with a being the lattice constant. Gray areas represent energy bands in which allowed states (dispersion curves) exist. States are occupied up to the Fermi level E_F as indicated by thick dispersion curves. (a) In insulators and semiconductors, all conduction band states are unoccupied at zero temperature and E_F lies in the energy gap. (b) In metals E_F lies in the conduction band and the conduction band is partially occupied resulting in finite conductivity.

1.3 Semiconducting materials

Semiconducting materials are numerous and versatile. We distinguish elementary and compound semiconductors.

Elementary semiconductors. Silicon (Si) and germanium (Ge), phosphorous (P), sulfur (S), selenium (Se), and tellurium (Te) are *elementary* semiconductors. Silicon is of utmost importance for the semiconductor industry. Certain modifications of carbon (C_{60}, nanotubes, graphene) can be called semiconductors.

Compound semiconductors. *Compound* semiconductors are classified according to the group of their constituents in the periodic table of elements (see Fig. 1.7). Gallium arsenide (GaAs), aluminium arsenide (AlAs), indium arsenide (InAs), indium antimonide (InSb), gallium antimonide (GaSb), gallium phosphide (GaP), gallium nitride (GaN), aluminium antimonide (AlSb), and indium phosphide (InP), for example, all belong to the so-called **III-V semiconductors**. In addition, there are **II-VI semiconductors**, such as zinc sulfide (ZnS), zinc selenide (ZnSe) and cadmium telluride (CdTe), **III-VI compounds**, such as gallium sulfide (GaS) and indium selenide (InSe), as well as **IV-VI compounds**, such as lead sulfide (PbS), lead telluride (PbTe), lead selenide (PbSe), germanium telluride (GeTe), tin selenide (SnSe), and tin telluride (SnTe). Among the more exotic semiconductor materials there are, for example, the copper oxides CuO and Cu_2O (cuprite), ZnO (zinc oxide), and PbS (lead sulfide, galena). Also of interest are **organic semiconductors** such as polyacetylene $(CH_2)_n$ or anthracene ($C_{14}H_{10}$).

I	II												III	IV	V	VI	VII	VIII
H																		He
Li	Be												B	C	N	O	F	Ne
Na	Mg												Al	Si	P	S	Cl	Ar
K	Ca		Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr		Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	*	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	**	Lr	Rf	Db	Sg	Bh	Hs	Mt									

Fig. 1.7 Periodic table of elements. Si and Ge in group IV, for example, are elementary semiconductors. Compound semiconductors contain, for example, elements from groups III and V, or II and VI.

Binary and ternary compounds. Compound semiconductors with two chemical constituents are called *binary compounds*. In addition, there are compound semiconductors with three constituents, such as $Al_xGa_{1-x}As$ (aluminium gallium arsenide), $In_xGa_{1-x}As$ (indium gallium arsenide), $In_xGa_{1-x}P$ (indium gallium phosphide), and also $CuFeS_2$ (chalcopyrite). In this case, one talks about *ternary semiconductors* or semiconductor alloys. They play an important role for the so-called 'bandgap engineering' which will be discussed in a later chapter.

In this book, with its focus on electronic transport in semiconductor nanostructures, the emphasis is often put on III-V semiconductors or on silicon. The reason is that there exists a very mature technology for fabricating nanostructures from these materials and because an extraordinary purity of these materials can be achieved. Both properties are extremely important for observing the quantum transport effects discussed later on.

Further reading

- Kittel 2005; Ashcroft and Mermin 1987; Singleton 2001; Seeger 2004; Cohen and Chelikowski 1989; Yu and Cardona 2001; Balkanski and Wallis 2000.
- Papers: Wilson 1931*a*; Wilson 1931*b*.

Exercises

(1.1) The 'Landolt–Börnstein' is an important series of data handbooks, also containing data about semiconductors. Find out where and how you have access to this reference. Find the volumes in which data about the semiconductors Si and GaAs can be found. Look up the values E_g of the band gaps of these two materials.

(1.2) You order a silicon wafer of 0.5 mm thickness and a resistivity of 10 Ωcm. What is the resistance of a bar of 1 cm width and 10 cm length, if measured

between the two ends of the bar? Compare the result to the resistance of a piece of copper having the same size. How much bigger is it?

(1.3) Find out which processor is used in your computer. Research on the internet how many transistors there are in the processor, and what the minimum pattern size is.

(1.4) Find all the Nobel prize winners who obtained their prize for important discoveries and/or contributions to modern semiconductor technology, and discuss their achievements.

(1.5) Assume that a single bit in an SRAM memory consisting of six transistors occupies a total area of $400\,\text{nm} \times 150\,\text{nm}$. What is the area needed for a 1 GB memory?

Semiconductor crystals

2

2.1 Crystal structure

Diamond and zincblende structure. Semiconductors form periodic crystal lattices. Silicon and germanium crystallize in the diamond lattice (see Fig. 2.1), whereas GaAs, AlAs, InAs, GaSb, for example, have a zincblende structure.

The diamond structure is an fcc lattice with a basis consisting of two atoms of the same kind (see Fig. 2.1). The zincblende lattice looks like the diamond lattice, but the two atoms forming the basis of the fcc lattice are different (e.g. Ga and As in GaAs, see Fig. 2.1).

Notation for crystal directions. Directions in a crystal are denoted in square brackets. The z-direction, for example, is described by [001]. Negative directions have a bar. For example, the $-z$-direction is $[00\bar{1}]$.

Notation for lattice planes: Miller indices. Lattice planes (all parallel planes) are labeled with the so-called Miller indices in round brackets. The normal vector characterizes the orientation of the plane. Integer numbers are chosen for the components of this vector. These are the Miller indices. The x-y plane, for example, is described by (001). Important orientations of crystal surfaces are the (001), the (111), the (110), the $(1\bar{1}0)$, and the (311) directions.

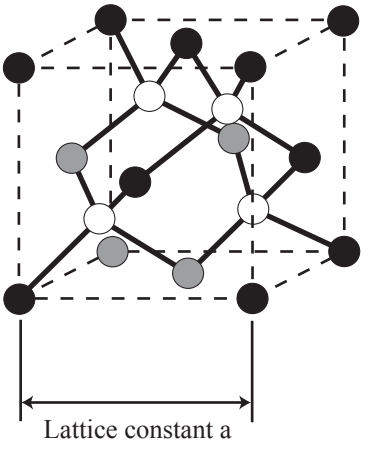


Fig. 2.1 Crystal structure of diamond. The spheres represent the positions of the atoms in the lattice. The zincblende structure is identical, but neighboring atoms are different elements (e.g. Ga and As).

2.2 Fabrication of crystals and wafers

2.2.1 Silicon

Reduction of silica. The fabrication of high purity silicon wafers from quartz sand for the semiconductor industry is depicted in Fig. 2.2 and briefly described below. The earth's crust contains a 25.7% by weight of silicon. There are enormous resources in the silicon dioxide (SiO_2, quartz, silica) contained in quartz sand. Silica makes the sand glitter in the sunlight. Silicon is made from silica in a furnace at 2000°C by reduction with carbon (coke) from the reaction

$$SiO_2 + 2C \rightarrow Si + 2CO.$$

This material has a purity of 97%.

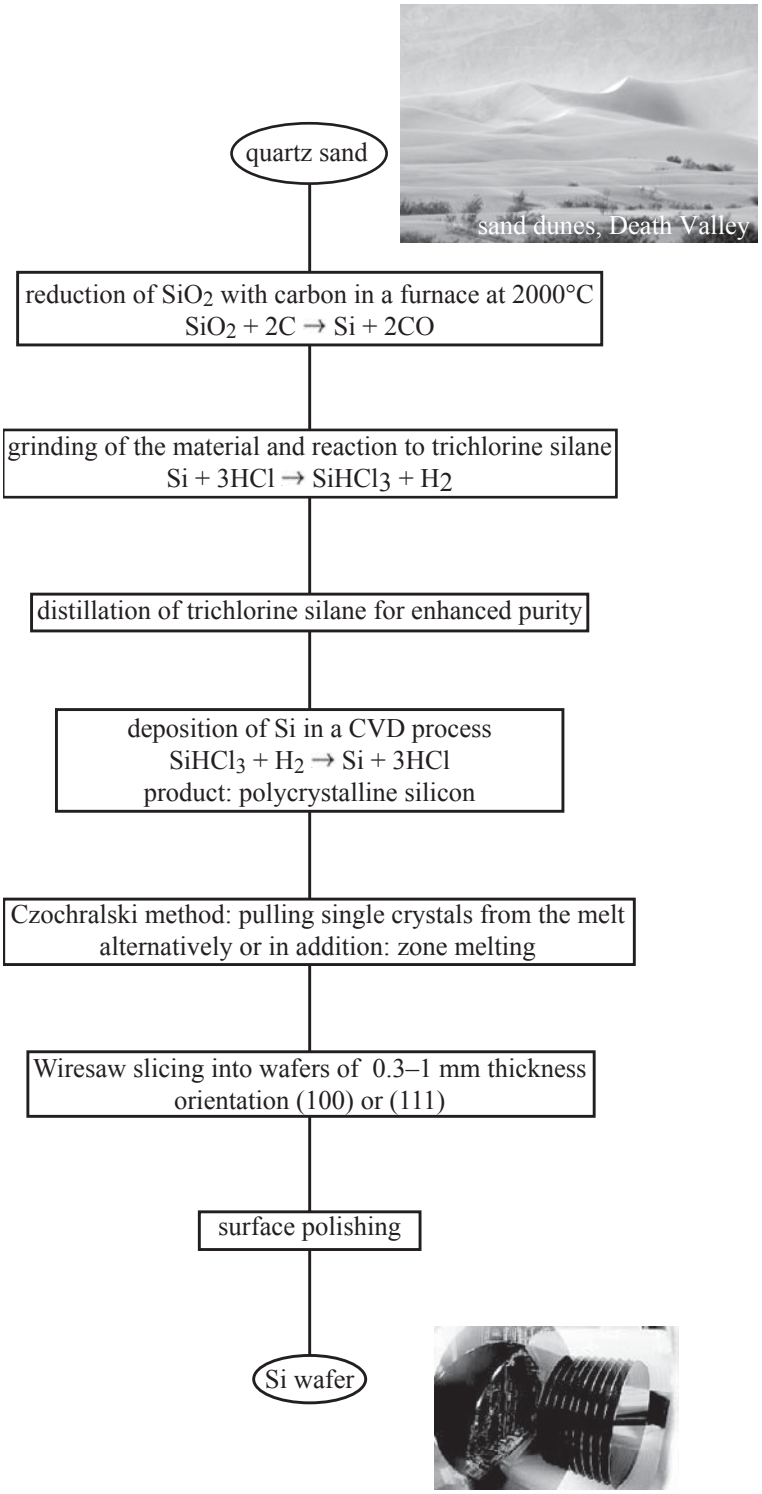


Fig. 2.2 Steps for the fabrication of high purity silicon wafers.

Chemical purification. The raw material is milled and mixed with hydrochloric acid (HCl). Under this influence it reacts to trichlorosilane ($SiHCl_3$) according to

$$Si + 3HCl \rightarrow SiHCl_3 + H_2$$

and impurities such as Fe, Al and B are removed. The purity of trichlorosilane can be increased by distillation. In a subsequent CVD (chemical vapor deposition) process, polycrystalline silicon is deposited containing less than 0.01 ppb of metallic impurities and less than 0.001 ppb of dopants (meaning 99.99999999% of Si):

$$SiHCl_3 + H_2 \rightarrow Si + 3HCl.$$

At this stage, doping atoms can be deliberately added.

Single crystal ingots. Large single crystals, so-called ingots, are then obtained by pulling the crystal from the melt (Si melts at 1420°C) of the polycrystalline material (Czochralski method, after the polish scientist J. Czochralski, 1916. See Fig. 2.3). Before this process, the chunks of polycrystalline material undergo thorough cleaning and surface etching in a cleanroom environment. Alternatively, single crystals are produced using zone melting, which is also an appropriate method for further cleaning existing single crystals. The end product is single crystals with a length of 1–2 m and a diameter of up to just over 30 cm (see Fig. 2.4). The density of dislocations in these single crystals is smaller than 1000 cm^{-3} (Yu and Cardona, 2001)[1], and the ratio of the number of impurity atoms to silicon atoms is smaller than 10^{-12}.

Grinding, slicing, and polishing. A mechanical rotary grinding process gives the ingot a perfect cylindrical shape. Wiresaw slicing normal to the cylinder axis produces flat silicon disks (so-called wafers) of about 0.3 mm to 1 mm thickness. The surfaces are typically in (100) or (111) direction and will be polished (by lapping and etching). On the basis of such silicon wafers, transistor circuits, including computer processors, can be fabricated.

2.2.2 Germanium

Germanium is extracted, like silicon, from its oxide, germanium dioxide (GeO_2) by reduction with carbon. High purity Germanium is obtained via $GeCl_4$, in analogy with the processes used for silicon. Large single crystals are pulled using the Czochralski method or zone melting. Natural germanium contains five different isotopes. Nowadays, germanium crystals can be made that contain only one particular isotope.

[1]Traditionally dislocation density is given per cm^2, because it is a density of line defects cut by a cross-section through the crystal. However, modern electron microscopy, or X-ray diffraction techniques give defect densities per cm^3 and thereby also capture bent dislocation defects that will not appear at the surface, e.g., of thin film samples (Yu, 2009).

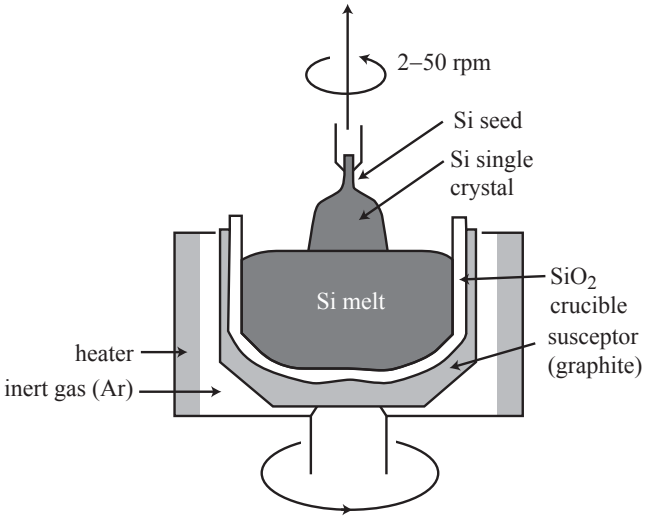


Fig. 2.3 Schematic of the Czochralski method for pulling semiconductor crystals from the melt (Yu and Cardona, 2001).

Fig. 2.4 Silicon single crystal, fabricated with the Czochralski method. The crystal has a diameter of 20 cm and a length of almost 2 m. It is suspended from the thin seed crystal (see upper right inset). (Copyright Kay Chernush, reproduced with permission).

2.2.3 Gallium arsenide

High pressure compounding. The compound III-V semiconductor gallium arsenide is fabricated from high purity gallium and arsenic. The exothermal reaction forming GaAs occurs at sufficiently high temperature and high pressure (compounding). Doping is possible during this step.

Single crystal ingots. Single crystals are pulled employing the Czochralski method. The GaAs melt is covered with liquid boron oxide (B_2O_3), in order to avoid the discharge of volatile anionic vapor. This is referred to as the LEC method (liquid-encapsulated Czochralski method). The quartz crucible can be used only once. It breaks when the remaining melt cools down. Alternatively, boron nitride crucibles can be used.

Compared to silicon, gallium arsenide single crystals cannot be purified very well. Silicon contaminants originate from the crucible and carbon from the graphite heaters and other parts of the apparatus. So-called semi-insulating GaAs is fabricated by compensating for shallow donors with deep acceptors (e.g., Si, Cr) and shallow acceptors with deep donors (e.g., C). If crucibles made of boron oxide are used, so-called undoped GaAs can be produced. The density of dislocations depends on the diameter of the crystal and is for two- or three-inch material of the order of $10^4 - 10^5\,\mathrm{cm}^{-2}$. The density of dislocations is typically smallest in the center of the single crystal.

Grinding, slicing and polishing. The pulled crystals are oriented and cut into thin wafers with two- or three-inch diameter and 0.015–0.035 in = 0.4–0.9 mm thickness. Surface polishing leads to wafer material that is ready for the fabrication of electronic devices.

2.3 Layer by layer growth

2.3.1 Molecular beam epitaxy – MBE

What is the meaning of 'epitaxy'? The word epitaxy consists of two ancient Greek words: first, *epi* (*επί*) means 'onto', and second, *taxis* (*τάξιζ*) means 'arranging' or 'ordering', but also the resulting 'arrangement'. The word expresses the process of growing additional crystal layers onto the surface of a substrate.

How it works. Starting from a semiconductor wafer, crystals can be grown with the so-called molecular beam epitaxy (MBE). One could call this method, which requires pressures of 10^{-10} to 10^{-11} mbar in the ultra high vacuum (UHV) regime, a refined evaporation technique. The wafer substrate is mounted in the UHV chamber on a substrate holder that can be heated (see Fig. 2.5). Atoms of different elements are evaporated from effusion cells that work like little ovens (Knudsen cells). The atom beams hit the heated substrate, atoms stick to the surface and

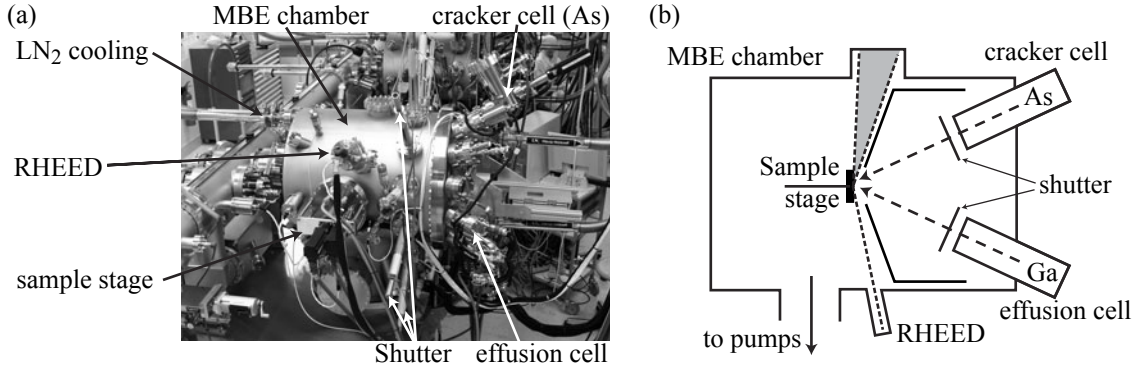


Fig. 2.5 (a) MBE system for arsenide epitaxy in the FIRST Center for Micro- and Nanoscience, ETH Zurich. The length of the chamber is roughly 1 m. (Image courtesy of H. Rusterholz and S. Schön.) (b) Schematic cross-section of an MBE-chamber.

diffuse around on the surface until they have found the energetically most favorable place in the crystal lattice. Typical growth temperatures are between 500°C and 600°C. Almost every material combination including doping can be grown, if the flux of the atoms (e.g., Ga, As, Al, Si, In) is controlled with shutters, and the substrate temperature is appropriate. In the right regime, the crystal grows atomic layer by atomic layer. In this way, very sharp transitions between materials (interfaces) and very sharp doping profiles can be achieved. A typical growth rate is one monolayer per second, or about 1 µm per hour.

In-situ observation of crystal growth. In-situ analysis of the crystal growth is facilitated by the fact that it takes place in UHV. Typically the RHEED (reflected high-energy electron diffraction) method is implemented. The method consists of scattering an electron beam incident under a very small angle at the surface [see Fig. 2.5(b)]. The resulting diffraction pattern is observed on a fluorescent screen. In the case of layer by layer growth, the RHEED intensity oscillates periodically, because the morphology of the surface changes periodically. This is a way of counting the number of atomic layers during growth.

Who operates MBE machines? MBE machines are operated by leading research labs and in industry. They grow, for example, Si, Ge, SiGe, GaAs/AlGaAs heterostructures and all kinds of other III-V or II-VI materials and heterostructures.

Which materials can be combined? In order to grow a certain layer sequence consisting of different materials, their lattice constants have to match reasonably well. For example, GaAs almost perfectly matches AlAs, as does the ternary alloy $Al_xGa_{1-x}As$. Extraordinary quality samples can be grown with this material system. Interfaces between the materials have a roughness of not more than one atomic layer. Such

layer sequences containing different materials are called *heterostructures*. They are an ideal starting point for the fabrication of more complicated semiconductor nanostructures.

Increasing substrate quality. Lattice dislocations in the substrate tend to propagate further into the growing crystal thereby impairing its quality. In the case of GaAs the material quality can be significantly improved by either growing a very thick GaAs layer on top of the substrate, or by repeatedly growing a few monolayers of GaAs and AlAs (short period superlattice). Also for other materials, such buffer layers were successfully employed.

Strained layers. If the lattice constants of subsequent layers are not perfectly matched, strain will develop in the crystal around the interface. The strain is typically released by the formation of lattice dislocations if the top layer grows beyond a certain critical thickness. Relatively thin layers, however, can be grown in a matrix of non-lattice-matched materials without the formation of dislocations. Such layers are called *pseudomorphic*.

Advantages of MBE. Using MBE, the growth of almost arbitrary materials is possible. A suitable sequence of layers leads to a layer quality that can be significantly improved over that of the substrate (e.g., fewer dislocations or impurities). In a good machine for GaAs, the background doping (i.e., the concentration of unintentionally incorporated impurity atoms) can be below 5×10^{13} cm^{-3}.

MBE machines allow us to control the layer thicknesses on the atomic scale, and also doping can be incorporated with atomic precision. Crystal growth is very homogeneous across the whole wafer, if the wafer is rotated.

Disadvantages compared to other methods. The main disadvantage of MBE machines is the cost of purchase and maintenance. The machines are also very complex and have very stringent vacuum requirements making involved and expensive pumping systems crucial.

2.3.2 Other methods

Other epitaxial methods are, for example, the 'vapor phase epitaxy' (VPE), the 'metal-organic chemical vapour deposition' (MOCVD) and the 'liquid phase epitaxy' (LPE). The MOCVD method is widely used and will therefore be briefly discussed below.

MOCVD Growing GaAs crystals with VPE brings the elements (e.g., Ga, As or doping atoms) in gaseous phase to the wafer surface. The MOCVD method is a variant of this principle, where gallium is supplied in the form of trimethyl gallium. The highly toxic AsH_3 gas is used as

the arsenic source. Aluminium can be supplied in the form of trimethyl aluminium. The main problems of this method are safety issues related to the toxic gases.

Further reading

- Crystal structure: Kittel 2005; Ashcroft and Mermin 1987; Singleton 2001; Yu and Cardona 2001.
- Fabrication of semiconductor crystals: Yu and Cardona 2001.

Exercises

(2.1) Given the lattice constant a, determine the following characteristic quantities for the simple cubic, body centered cubic (bcc), face centered cubic (bcc), and diamond lattices: (a) unit cell volume, (b) number of atoms in the unit cell, (c) primitive cell volume, (d) coordination number, (e) nearest neighbor separation.

(2.2) The density of silicon is $\rho_{\mathrm{Si}} = 2330\,\mathrm{kg/m^3}$. Calculate the side length of the cubic unit cell and the separation of neighboring silicon atoms.

(2.3) Find points in the unit cell of silicon that are symmetry points with respect to spatial inversion. Spatial inversion around the origin of the coordinate system transforms a vector (x, y, z) into $(-x, -y, -z)$.

(2.4) Does the GaAs crystal have points of inversion symmetry? Explain.

(2.5) A silicon wafer with a thickness $t = 200\,\mu\mathrm{m}$ has an initial weight $m_0 = 46.6\,\mathrm{mg}$. After thermal oxidation forming an SiO_2 covered surface, the same wafer has increased its weight to $m_1 = 46.89\,\mathrm{mg}$. The density of silicon is $\rho_{\mathrm{Si}} = 2.33\,\mathrm{g/cm^3}$ and that of the oxide is $\rho_{\mathrm{oxide}} = 2.20\,\mathrm{g/cm^3}$. Determine the thickness of the oxide layer and the reduction in thickness of the pure silicon material.

(2.6) The UHV chamber of an MBE machine has a diameter of the order of 1 m. Estimate the pressure required in the chamber for atoms to traverse it ballistically, i.e. without collisions.

(2.7) Estimate the rate at which gas molecules of mass m in a gas with pressure p at temperature T hit the surface of a substrate.

(2.8) Estimate how long it takes in an MBE chamber for a monolayer of oxygen atoms to form at the substrate surface, given that the background gas is at room temperature and has a partial oxygen pressure of $10^{-10}\,\mathrm{mbar}$. Assume that all impinging atoms stick to the surface and use the kinetic theory of gases.

(2.9) Estimate the required growth rate in an MBE chamber with a background pressure of $10^{-10}\,\mathrm{mbar}$ which makes sure that less than $10^6\,\mathrm{cm^{-2}}$ impurities are incorporated in a single atomic plane of the crystal.

Band structure

3

3.1 Spinless and noninteracting electrons

The basic problem. The band structure of semiconductors emerges as a solution of Schrödinger's equation for noninteracting electrons in the periodic potential of the crystal lattice:

$$\left[-\frac{\hbar^2}{2m_\mathrm{e}}\Delta + V(\mathbf{r})\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}), \tag{3.1}$$

where the potential has the property

$$V(\mathbf{r}) = V(\mathbf{r}+\mathbf{R}). \tag{3.2}$$

The vector $\mathbf{R}$ is an arbitrary translation vector that moves the lattice onto itself.

Fourier expansion of the potential and reciprocal lattice. Owing to its periodicity, the crystal potential can be expanded in a Fourier series:

$$V(\mathbf{r}) = \sum_{\mathbf{G}} V_{\mathbf{G}} e^{i\mathbf{G}\mathbf{r}}. \tag{3.3}$$

The allowed vectors of the reciprocal lattice $\mathbf{G}$ are determined from the periodicity of the lattice, eq. (3.2):

$$V(\mathbf{r}) = \sum_{\mathbf{G}} V_{\mathbf{G}} e^{i\mathbf{G}\mathbf{r}} = \sum_{\mathbf{G}} V_{\mathbf{G}} e^{i\mathbf{G}(\mathbf{r}+\mathbf{R})} = \sum_{\mathbf{G}} V_{\mathbf{G}} e^{i\mathbf{G}\mathbf{r}} e^{i\mathbf{G}\mathbf{R}} \overset{(!)}{=} V(\mathbf{r}+\mathbf{R}).$$

This gives the condition

$$e^{i\mathbf{G}\mathbf{R}} = 1, \text{ or } \mathbf{G}\mathbf{R} = 2\pi n,$$

where n is an integer number.

The reciprocal lattice of an fcc lattice with lattice constant a is a bcc lattice with lattice constant $2\pi/a$. Table 3.1 shows the shortest reciprocal lattice vectors of an fcc lattice.

First Brillouin zone. The first Brillouin zone comprises those points in reciprocal lattice space that are closer to the origin (i.e., to the Γ point) than to any other point of the reciprocal lattice. As an example, Fig. 3.1 shows the first Brillouin zone of the fcc lattice. Points of high symmetry are commonly labeled with capital letters Γ, X, L, U, K, W. Their coordinates are given in Table 3.2.

Table 3.1 The shortest reciprocal lattice vectors $\mathbf{G}'$ of an fcc lattice. Lengths are in units of $2\pi/a$.

Origin: $(0,0,0)$
Nearest neighbors: $\pm(1,1,1)$ $\pm(-1,1,1)$ $\pm(1,-1,1)$ $\pm(-1,-1,1)$
Next nearest neighbors: $\pm(2,0,0)$ $\pm(0,2,0)$ $\pm(0,0,2)$
$\pm(2,2,0)$ $\pm(2,-2,0)$ $\pm(2,0,2)$ $\pm(2,0,-2)$ $\pm(0,2,2)$ $\pm(0,2,-2)$

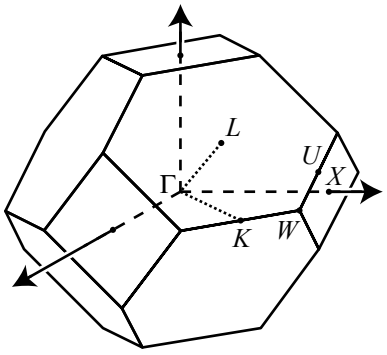


Fig. 3.1 First Brillouin zone of the fcc lattice. The points Γ, X, L, and others are indicated.

Table 3.2 Coordinates of symmetry points in the reciprocal lattice of an fcc lattice. Lengths are in units of $2\pi/a$.

Γ	$(0,0,0)$
X	$(1,0,0)$
L	$(1/2,-1/2,1/2)$
U	$(1,-1/4,1/4)$
K	$(3/4,-3/4,0)$
W	$(1,\sqrt{2}/2,0)$

Band structure equation and Bloch's theorem. With the Fourier expansion of the potential, eq. (3.3), Schrödinger's equation (3.1) reads

$$\left[-\frac{\hbar^2}{2m_\mathrm{e}}\Delta+\sum_{\mathbf{G}}V_{\mathbf{G}}e^{i\mathbf{G}\mathbf{r}}\right]\psi(\mathbf{r})=E\psi(\mathbf{r}), \tag{3.4}$$

This differential equation can be transformed into an algebraic equation by expanding the wave functions $\psi(\mathbf{r})$ in the Fourier series

$$\psi(\mathbf{r})=\sum_{\mathbf{q}}c_{\mathbf{q}}e^{i\mathbf{q}\mathbf{r}}. \tag{3.5}$$

The values of $\mathbf{q}$ are, for example, restricted by the assumption of periodic boundary conditions (Born–von Karman boundary conditions). However, the values of $\mathbf{q}$ are so dense, owing to the macroscopic size of the crystal, that we can regard this vector as being quasi-continuous. Inserting this expansion into Schrödinger's equation (3.4) gives

$$\sum_{\mathbf{q}}e^{i\mathbf{q}\mathbf{r}}\left[\left(\frac{\hbar^2\mathbf{q}^2}{2m_\mathrm{e}}-E\right)c_{\mathbf{q}}+\sum_{\mathbf{G}}V_{\mathbf{G}}c_{\mathbf{q}-\mathbf{G}}\right]=0.$$

Multiplying this equation by $e^{-i\mathbf{q}'\mathbf{r}}$ and integrating over $\mathbf{r}$ we see that each Fourier component obeys the equation

$$\left(\frac{\hbar^2\mathbf{q}^2}{2m_\mathrm{e}}-E(\mathbf{q})\right)c_{\mathbf{q}}+\sum_{\mathbf{G}}V_{\mathbf{G}}c_{\mathbf{q}-\mathbf{G}}=0. \tag{3.6}$$

Here, we have introduced $E(\mathbf{q})\equiv E$ for denoting the quasi-continuous energy dispersion depending on the wave vector $\mathbf{q}$. An arbitrary vector $\mathbf{q}$ can be mapped on a vector $\mathbf{k}$ in the first Brillouin zone by adding a suitable reciprocal lattice vector $\mathbf{G}'$, i.e., $\mathbf{k}=\mathbf{q}+\mathbf{G}'$. With this notation we find from eq. (3.6):

$$\left(\frac{\hbar^2(\mathbf{k}-\mathbf{G}')^2}{2m_\mathrm{e}}-E(\mathbf{k})\right)c_{\mathbf{k}-\mathbf{G}'}+\sum_{\mathbf{G}}V_{\mathbf{G}}c_{\mathbf{k}-\mathbf{G}'-\mathbf{G}}=0.$$

Since $\mathbf{G}+\mathbf{G}'$ is itself a reciprocal lattice vector, we introduce $\mathbf{G}''=\mathbf{G}+\mathbf{G}'$ and obtain

$$\left(\frac{\hbar^2(\mathbf{k}-\mathbf{G}')^2}{2m_\mathrm{e}}-E(\mathbf{k})\right)c_{\mathbf{k}-\mathbf{G}'}+\sum_{\mathbf{G}''}V_{\mathbf{G}''-\mathbf{G}'}c_{\mathbf{k}-\mathbf{G}''}=0. \tag{3.7}$$

This is the desired algebraic equation for the coefficients $c_{\mathbf{k}-\mathbf{G}'}$ and the energies $E(\mathbf{k})$. For any given vector $\mathbf{G}'$ a particular dispersion relation $E_{\mathbf{G}'}(\mathbf{k})$ results. We can introduce a band index n replacing this vector, because the lattice of possible vectors $\mathbf{G}'$ is discrete. Then we talk about the nth energy band with dispersion relation $E_n(\mathbf{k})$. Eq. (3.7) is thereby the equation for determining the band structure of a solid.

Equation (3.7) contains only coefficients $c_{\mathbf{q}}$ of the wave function (3.5) in which $\mathbf{q}=\mathbf{k}-\mathbf{G}$, with $\mathbf{G}$ being a reciprocal lattice vector. Therefore,

for given $\mathbf{k}$, there is a wave function $\psi_{\mathbf{k}}(\mathbf{r})$ that solves Schrödinger's equation and takes the form

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} e^{i(\mathbf{k}-\mathbf{G})\mathbf{r}} = e^{i\mathbf{k}\mathbf{r}} \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} e^{-i\mathbf{G}\mathbf{r}} := e^{i\mathbf{k}\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) \quad (3.8)$$

Here we have introduced functions $u_{n\mathbf{k}}(\mathbf{r})$ with the property

$$u_{n\mathbf{k}}(\mathbf{r}+\mathbf{R}) = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} e^{-i\mathbf{G}(\mathbf{r}+\mathbf{R})} = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} e^{-i\mathbf{G}\mathbf{r}} \underbrace{e^{-i\mathbf{G}\mathbf{R}}}_{=1} = u_{n\mathbf{k}}(\mathbf{r}). \quad (3.9)$$

The vector $\mathbf{R}$ is a translation vector of the crystal lattice. The function $u_{n\mathbf{k}}(\mathbf{r})$ has the translational symmetry of the lattice. The two eqs (3.8) and (3.9) express what is known as *Bloch's theorem*.

Pseudopotential method. The plane wave expansion shown above provides a straightforward formal way to calculate band structures. In practice, however, the problem arises that very large numbers of plane wave coefficients are significant which makes it hard to achieve numerical convergence taking only a reasonable number of states into account. Therefore, more refined methods make use of the fact that the inner shells of the atoms in a lattice are tightly bound. They are hardly influenced by the presence of the neighboring atoms. These *core states* can therefore be assumed to be known from the calculation of the atomic energy spectra.

The remaining task of calculating the extended states of the valence electrons can be simplified by constructing states that are orthogonal to the core states. In effect, the valence electrons are found to move in an effective potential (the so-called *pseudopotential*) which is the sum of the bare potential created by the nuclei and a contribution created by the orthogonality requirement to the core states. It can be shown that the energy levels of the valence and conduction band states can be obtained by solving the Schrödinger equation (3.7) containing the pseudopotential as a weak perturbation of free electron motion.

Although the pseudopotential method converges with a relatively small number of plane wave contributions, the problem remains to determine the (usually nonlocal) pseudopotential. In practice, the simplest solution is the use of empirical (often local) pseudopotentials that depend on parameters that can be adjusted such that the resulting band structure fits the results of measurements.

Free electron model. We obtain the lowest order approximation to the valence and conduction band structure of a semiconductor by completely neglecting the lattice periodic (pseudo)potential contribution in eq. (3.7). The dispersion relation for a particular type of lattice is then given by

$$E_n(\mathbf{k}) = \frac{\hbar^2(\mathbf{k}-\mathbf{G}')^2}{2m_{\mathrm{e}}}. \quad (3.10)$$

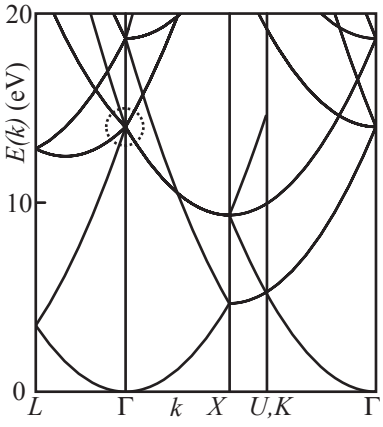


Fig. 3.2 Band structure of an fcc lattice in the free electron model.

As an example, we consider an fcc lattice. The reciprocal lattice is bcc and the shortest reciprocal lattice vectors are listed in Table 3.1. Fig. 3.2 shows the resulting band structure along certain straight lines connecting symmetry points in the first Brillouin zone. Degeneracies occur at points Γ, X, L, U, and K (cf. Fig. 3.1) whose coordinates are listed in Table 3.2. For example at L, the two parabolic dispersions coincide which have minima at $(0,0,0)$ and at $2\pi/a(1,1,1)$. An eight-fold degeneracy exists at the Γ-point in Fig. 3.2 (encircled) resulting from parabolae with minima at the nearest neighbors (Table 3.1). This degeneracy will be lifted leading to the band gap, and separate valence and conduction bands, if the lattice periodic potential is taken into account.

Pseudopotential method for diamond and zincblende semiconductors: a case study. The weak potential modulation acts strongest at degeneracy points of the free electron dispersion and tends to lift degeneracies at least partially. As a result, a band gap, i.e., an energetic region in which no states exist, will open up between valence and conduction bands.

In order to see this effect, matrix elements $V_{\mathbf{G}''-\mathbf{G}'}$ of the pseudopotential in eq. (3.7) will be required. The contributions with $\mathbf{G}'' = \mathbf{G}'$ lead to diagonal matrix elements V_0 that simply shift the dispersion curves in energy. Off-diagonal elements involving finite length reciprocal lattice vectors $\mathbf{G} = \mathbf{G}'' - \mathbf{G}'$ give significant contributions only for the shortest vectors.

As an example, we briefly discuss the pseudopotential method for diamond and zincblende structures. The Fourier transform of the lattice potential is

$$V_{\mathbf{G}} = \frac{1}{\Omega}\int_{\mathrm{PC}} d^3r\; V(\mathbf{r})e^{-i\mathbf{G}\mathbf{r}},$$

where the integration is performed over the primitive cell (PC) with volume Ω. In diamond or zincblende crystals the PC contains two atoms A and B and we write the pseudopotential as the sum of two atomic pseudopotentials, i.e., $V(\mathbf{r}) = V_{\mathrm{A}}(\mathbf{r}-\mathbf{r}_{\mathrm{A}}) + V_{\mathrm{B}}(\mathbf{r}-\mathbf{r}_{\mathrm{B}})$. If we choose the origin at the midpoint between atoms A and B, we have $\mathbf{r}_{\mathrm{A}} = -\mathbf{r}_{\mathrm{B}} = a(1/8,1/8,1/8)$. As a consequence,

$$\begin{aligned} V_{\mathbf{G}} &= \frac{1}{\Omega}\int_{\mathrm{PC}} d^3r\; V_{\mathrm{A}}(\mathbf{r}-\mathbf{r}_{\mathrm{A}})e^{-i\mathbf{G}\mathbf{r}} + \frac{1}{\Omega}\int_{\mathrm{PC}} d^3r\; V_{\mathrm{B}}(\mathbf{r}-\mathbf{r}_{\mathrm{B}})e^{-i\mathbf{G}\mathbf{r}} \\ &= e^{-i\mathbf{G}\mathbf{r}_{\mathrm{A}}}\frac{1}{\Omega}\int_{\mathrm{PC}} d^3r\; V_{\mathrm{A}}(\mathbf{r})e^{-i\mathbf{G}\mathbf{r}} + e^{-i\mathbf{G}\mathbf{r}_{\mathrm{B}}}\frac{1}{\Omega}\int_{\mathrm{PC}} d^3r\; V_{\mathrm{B}}(\mathbf{r})e^{-i\mathbf{G}\mathbf{r}}. \end{aligned}$$

We see that the Fourier transforms of the pseudopotentials of atoms A and B enter as parameters. They depend only on $|\mathbf{G}|$ owing to the symmetry of the core electronic states. The exponential prefactors are called *structure factors*. Defining

$$V_G^{\mathrm{A/B}} = \frac{1}{\Omega}\int_{\mathrm{PC}} d^3r\; V_{\mathrm{A/B}}(\mathbf{r})e^{-i\mathbf{G}\mathbf{r}},$$

we can write the matrix elements as

$$V_{\mathbf{G}} = \underbrace{(V_G^{\mathrm{A}} + V_G^{\mathrm{B}})}_{V_G^{\mathrm{s}}} \cos(\mathbf{G}\mathbf{r}_{\mathrm{A}}) - i \underbrace{(V_G^{\mathrm{A}} - V_G^{\mathrm{B}})}_{V_G^{\mathrm{a}}} \sin(\mathbf{G}\mathbf{r}_{\mathrm{A}}).$$

As a consequence, only symmetric (V_G^{s}) and asymmetric (V_G^{a}) combinations of $V_G^{\mathrm{A/B}}$ enter into the calculation.

The particular symmetries of the lattices leads to considerable simplifications. For diamond lattices, atoms A and B are identical, and therefore $V_G^{\mathrm{a}} = 0$, i.e., all matrix elements are real. The diagonal matrix element $V_{\mathbf{G}=0} = V_0^{\mathrm{s}}$ is always real and leads to an overall energy shift, as mentioned above. Matrix elements $V_{|\mathbf{G}|^2=4} \equiv \pm i V_4$ are purely imaginary for zincblende semiconductors. Matrix elements $V_{|\mathbf{G}|^2=8} \equiv \pm V_8$ are real also for zincblende crystals.

Figure 3.3 shows the result of such a pseudopotential calculation for silicon which can nowadays easily be implemented on a standard personal computer. The 51×51 hamiltonian matrix was diagonalized numerically in Mathematica using the three empirical pseudopotential parameters $V_3^{\mathrm{s}} = -2.87$ eV, $V_8^{\mathrm{s}} = 0.544$ eV, and $V_{11}^{\mathrm{a}} = 1.09$ eV. Matrix elements for longer reciprocal lattice vectors were set to zero. The zero of energy was chosen to be the valence band maximum at Γ. Comparison with the free electron model in Fig. 3.2 shows many similarities. However, pronounced gaps have opened, for example, at L and Γ. The fundamental band gap in silicon (shaded in gray) is between the valence band maximum at Γ and the conduction band minimum near X.

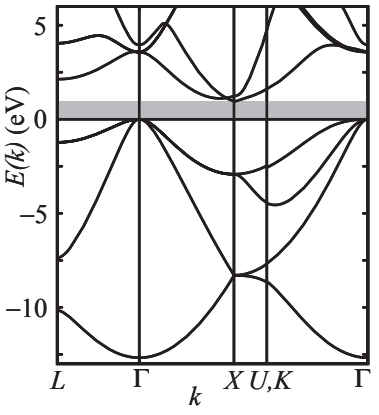


Fig. 3.3 Result of local pseudopotential calculations for silicon. The fundamental band gap is shaded in gray.

Figure 3.4 is the result of a similar calculation performed for GaAs having different A and B atoms in the primitive cell. In this case, five nonzero parameters are necessary due to the finite asymmetric contributions. The parameters used in this calculation were $V_3^{\mathrm{s}} = -3.43$ eV, $V_4^{\mathrm{s}} = V_8^{\mathrm{s}} = 0$, $V_{11}^{\mathrm{s}} = 1.09$ eV, $V_3^{\mathrm{a}} = 0.925$ eV, $V_4^{\mathrm{a}} = 0.90$ eV, and $V_{11}^{\mathrm{a}} = 0.163$ eV. Unlike in silicon, in GaAs the fundamental band gap appears between the valence band maximum at Γ and the conduction band minimum at Γ. The conduction band minima at L and X are higher in energy.

Better approximations, beyond the presented empirical pseudopotential method, take nonlocal pseudopotentials into account, sometimes even including interaction effects self-consistently. As discussed in the next section, an important ingredient missing so far for determining the band structure is the spin–orbit interaction.

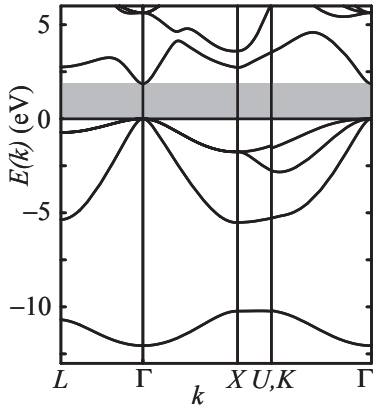


Fig. 3.4 Result of local pseudopotential calculations for GaAs. Spin–orbit interaction effects were neglected. The fundamental band gap is shaded in gray.

Tight-binding approximation. So far we have discussed band structure calculations using the strategy of the plane-wave expansion (3.5). In some cases, a different approach called the tight-binding approximation, leads to useful results. It regards the atoms in the lattice as weakly interacting, such that the atomic orbitals remain (almost) intact. The wave function for electrons in a particular band is a linear combination of degenerate wave functions that are not too different from atomic wave functions. The linear combination is chosen such that the wave function fulfills Bloch's theorem (Ashcroft and Mermin, 1987).

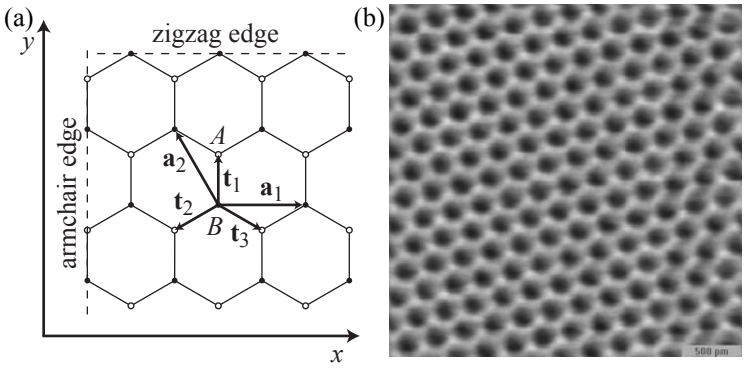


Fig. 3.5 (a) Graphene has a hexagonal lattice with a two-atom basis (atoms A and B). Lattice vectors are $\mathbf{a}_1$ and $\mathbf{a}_2$. Characteristic edges are also indicated. The armchair edge has A and B atoms, whereas the zigzag edge has only one type of atom (A or B). (b) Scanning tunneling microscopy image of the graphene lattice with atomic resolution showing the hexagonal lattice structure (Li, 2003. Image courtesy of Eva Andrei, Rutgers University).

The best-known example is the approximate calculation of the band structure of a single layer of graphite called *graphene*. Graphene is a two-dimensional atomic plane of carbon atoms that are arranged in a planar honeycomb lattice as shown in Fig. 3.5(a) and (b) with a bond length of about 0.14 nm. The better known graphite consists of stacked graphene planes. Graphene is an interesting material in current research: Although its band structure had already been calculated in 1947 (Wallace), single layer graphene sheets have only recently become available for condensed matter research (Novoselov *et al.*, 2004). They exhibit a fascinating variant of the quantum Hall effect (Novoselov *et al.*, 2005; Zhang *et al.*, 2005). Graphene sheets can also be rolled up to form carbon nanotubes which are also fascinating research objects in mesoscopic physics and other fields. Band structure calculations for graphene are facilitated compared to those of three-dimensional crystals, because of the two-dimensionality of the problem. We show the details of the calculation of the graphene π- and $\pi^\star$ bands in order to illustrate the method of using a linear combination of atomic orbitals.

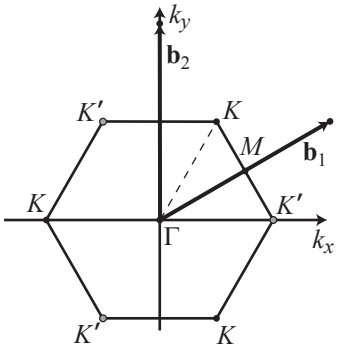


Fig. 3.6 First Brillouin zone of graphene with symmetry points Γ, K, K', and M, and the reciprocal lattice vectors $\mathbf{b}_1$ and $\mathbf{b}_2$.

In graphene, the carbon atoms are bound via sp^2-hybrid orbitals forming σ-bonds in the plane of the hexagonal lattice. Each carbon atom contributes three of its four valence electrons to σ-bonds. The fourth valence electron occupies the p_z-orbital. Overlapping p_z-orbitals form π-bonds between neighboring atoms, or—in the language of band structure—the π- and $\pi^\star$-bands.

The crystal lattice of graphene can be described as a Bravais lattice with two basis atoms A and B [see Fig. 3.5(a)]. For the Bravais lattice we choose basis vectors $\mathbf{a}_1 = a(1, 0)$ and $\mathbf{a}_2 = a(-1/2, \sqrt{3}/2)$, where $a = 2.46\,\text{Å}$. We take atoms B to sit at the sites $\mathbf{R} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2$ and atoms A at $\mathbf{R} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + \mathbf{t}_1$, with vectors $\mathbf{t}_1 = a(0, 1/\sqrt{3})$, $\mathbf{t}_2 = a(-1/2, -1/2\sqrt{3})$, and $\mathbf{t}_3 = a(1/2, -1/2\sqrt{3})$ pointing from the B atom at the origin to the nearest neighbor A atoms. The reciprocal lattice vectors are given by (see Fig. 3.6) $\mathbf{b}_1 = 2\pi/a(1, 1/\sqrt{3})$ and $\mathbf{b}_2 = 2\pi/a(0, 2/\sqrt{3})$. Points of high symmetry in the first Brillouin zone are the Γ-point, the points $\mathbf{K}$ and $\mathbf{K}'$, and the $\mathbf{M}$-points (see Table 3.3).

The wavefunction is taken to be the linear combination of atomic

orbitals

$$\psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{N}} \sum_{\mathbf{R}} e^{i\mathbf{k}\mathbf{R}} \left[A\phi(\mathbf{r} - \mathbf{R} - \mathbf{t}_1) + B\phi(\mathbf{r} - \mathbf{R}) \right], \tag{3.11}$$

where A and B are unknown amplitude parameters to be determined and N is the number of lattice sites in the crystal. The wave function $\phi(\mathbf{r})$ describes the p_z-orbital of the sp^2-hybridized carbon atom. The wavefunction (3.11) can also be written in the form of eq. (3.8)

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}), \tag{3.12}$$

with

$$u_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{N}} \sum_{\mathbf{R}} e^{-i\mathbf{k}(\mathbf{r}-\mathbf{R})} \left[A\phi(\mathbf{r} - \mathbf{R} - \mathbf{t}_1) + B\phi(\mathbf{r} - \mathbf{R}) \right]$$

having the periodicity of the crystal lattice. It therefore fulfills Bloch's theorem expressed in eqs (3.8) and (3.9).

The Hamiltonian for the crystal lattice is given by

$$H = \frac{\mathbf{p}^2}{2m_{\mathrm{e}}} + \sum_{\mathbf{R}} \left[V_0(\mathbf{r} - \mathbf{R} - \mathbf{t}_1) + V_0(\mathbf{r} - \mathbf{R}) \right].$$

Before we tackle the full eigenvalue problem with our trial wave function, we apply H to $\phi(\mathbf{r})$ and find

$$\begin{aligned} H\phi(\mathbf{r}) = &\left[\frac{\mathbf{p}^2}{2m_{\mathrm{e}}} + V_0(\mathbf{r}) \right] \phi(\mathbf{r}) \\ &+ V_0(\mathbf{r} - \mathbf{t}_1)\phi(\mathbf{r}) + \sum_{\mathbf{R}\neq 0} \left[V_0(\mathbf{r} - \mathbf{R} - \mathbf{t}_1) + V_0(\mathbf{r} - \mathbf{R}) \right] \phi(\mathbf{r}) \\ &:= \epsilon\phi(\mathbf{r}) + \Delta V_{\mathrm{B}}\phi(\mathbf{r}), \end{aligned}$$

where ϵ is the energy of the p_z-orbital in the carbon atom. Because we are free to set the zero of energy, we choose $\epsilon = 0$ and therefore have

$$H\phi(\mathbf{r}) = \Delta V_{\mathrm{B}}\phi(\mathbf{r}).$$

The right-hand side of this equation is small, because where ΔV_{B} is appreciable, $\phi(\mathbf{r})$ is small, and vice versa. Correspondingly,

$$H\phi(\mathbf{r} - \mathbf{t}_1) = \Delta V_{\mathrm{A}}\phi(\mathbf{r} - \mathbf{t}_1),$$

with small right-hand side and

$$\Delta V_{\mathrm{A}} = V_0(\mathbf{r}) + \sum_{\mathbf{R}\neq 0} \left[V_0(\mathbf{r} - \mathbf{R} - \mathbf{t}_1) + V_0(\mathbf{r} - \mathbf{R}) \right].$$

We are now ready to solve the eigenvalue problem $H\psi_{\mathbf{k}}(\mathbf{r}) = E\psi_{\mathbf{k}}(\mathbf{r})$ using the wave function (3.11). We solve this problem by projecting onto

Table 3.3 Coordinates of symmetry points in the reciprocal lattice of graphene. Lengths are in units of $2\pi/a$.

$\mathbf{\Gamma}$	$(0, 0)$
$\mathbf{K}$	$(1/3, 1/\sqrt{3})$
	$(1/3, -1/\sqrt{3})$
	$(-2/3, 0)$
$\mathbf{K}'$	$(2/3, 0)$
	$(-1/3, -1/\sqrt{3})$
	$(-1/3, 1/\sqrt{3})$
$\mathbf{M}$	$(1/2, 1/2\sqrt{3})$
	$(0, 1/\sqrt{3})$
	$(0, -1/\sqrt{3})$
	$(1/2, -1/2\sqrt{3})$
	$(-1/2, -1/2\sqrt{3})$
	$(-1/2, 1/2\sqrt{3})$

the two states $\phi(\mathbf{r})$ and $\phi(\mathbf{r}-\mathbf{t}_1)$. The projection leads to two equations for A and B, namely,

$$\begin{aligned}\int d^3r\phi^\star(\mathbf{r})H\psi_\mathbf{k}(\mathbf{r}) &= E\int d^3r\phi(\mathbf{r})^\star\psi_\mathbf{k}(\mathbf{r})\\ \int d^3r\phi^\star(\mathbf{r}-\mathbf{t}_1)H\psi_\mathbf{k}(\mathbf{r}) &= E\int d^3r\phi^\star(\mathbf{r}-\mathbf{t}_1)\psi_\mathbf{k}(\mathbf{r}).\end{aligned}$$

They can be transformed into

$$\begin{aligned}\left[\int d^3r\psi^\star_\mathbf{k}(\mathbf{r})\Delta V_\mathrm{B}\phi(\mathbf{r})\right]^\star &= E\int d^3r\phi(\mathbf{r})^\star\psi_\mathbf{k}(\mathbf{r})\\ \left[\int d^3r\psi^\star_\mathbf{k}(\mathbf{r})\Delta V_\mathrm{A}\phi(\mathbf{r}-\mathbf{t}_1)\right]^\star &= E\int d^3r\phi^\star(\mathbf{r}-\mathbf{t}_1)\psi_\mathbf{k}(\mathbf{r}).\end{aligned}$$

We now approximate the overlap integrals by considering only the contributions of nearest neighbor atoms. After some algebra this gives the matrix equation

$$\begin{pmatrix}\sigma-E & \alpha^\star(\mathbf{k})(\gamma-Es)\\ \alpha(\mathbf{k})(\gamma-Es) & \sigma-E\end{pmatrix}\begin{pmatrix}A\\B\end{pmatrix}=0 \qquad (3.13)$$

where

$$\begin{aligned}\alpha(\mathbf{k}) &= 1+e^{i\mathbf{k}(\mathbf{t}_2-\mathbf{t}_1)}+e^{i\mathbf{k}(\mathbf{t}_3-\mathbf{t}_1)}\\ \gamma &= \int d^3r\phi(\mathbf{r}-\mathbf{t}_1)V_0(\mathbf{r}-\mathbf{t}_1)\phi^\star(\mathbf{r})\\ \sigma &= 3\int d^3r\phi^\star(\mathbf{r})V_0(\mathbf{r}-\mathbf{t}_1)\phi(\mathbf{r})\\ s &= \int d^3r\phi^\star(\mathbf{r})\phi(\mathbf{r}-\mathbf{t}_1).\end{aligned}$$

We note here that the wave functions for the p_z-orbital can be chosen to be real and we have therefore applied $\gamma=\gamma^\star$, and $s=s^\star$. We find the energy eigenvalues from the characteristic equation

$$|\alpha(\mathbf{k})|^2(\gamma-Es)^2=(\sigma-E)^2.$$

As a further approximation, we neglect terms multiplying E that are second order in the overlap integrals on the left-hand side

$$(\gamma-Es)^2=\gamma^2-2\gamma sE+s^2E^2\approx\gamma^2,$$

because on the right-hand side, there are lower order terms. This leads to the two branches of the energy dispersion relation

$$\begin{aligned}&E(\mathbf{k})=\sigma\pm\gamma|\alpha(\mathbf{k})|\\ &=\sigma\pm\gamma\sqrt{3+2\cos[\mathbf{k}(\mathbf{t}_2-\mathbf{t}_1)]+2\cos[\mathbf{k}(\mathbf{t}_3-\mathbf{t}_1)]+2\cos[\mathbf{k}(\mathbf{t}_3-\mathbf{t}_2)]}.\end{aligned}$$

The constant σ merely gives an energy offset that is due to the first order energy shift of the atomic level under the influence of the remaining

crystal lattice. We redefine our energy offset accordingly and obtain the final result

$$E(\mathbf{k}) = \pm\gamma\sqrt{1 + 4\cos^2[k_x a/2] + 4\cos[k_x a/2]\cos[\sqrt{3}k_y a/2]},$$

where we have used the addition theorems for the trigonometric functions and the definitions of $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_3$ in order to simplify the expression. Figure 3.7 shows this dispersion relation describing the π-band, i.e., the valence band of graphene, and the $\pi^\star$-band which forms the conduction band along a specific line within the first Brillouin zone. A three-dimensional version of the dispersion is shown in Fig. 3.8. Most remarkably, a degeneracy of the dispersion remains at the **K**- and **K**′-points, while a gap opens at M. The π-band, which is lower in energy, takes two electrons per **k**-point (as a result of spin degeneracy). Therefore, the π-band is completely filled at zero temperature with the two p_z-electrons of the two atoms A and B forming the basis of the lattice. In contrast, the $\pi^\star$-band is completely empty. Graphene is sometimes referred to as a zero-gap semiconductor, because the zero temperature Fermi energy lies at the energy of the **K**- and **K**′-points. The linear dispersion at the **K**- and **K**′-points is one of the reasons why the appearance of some effects of mesoscopic physics is very different in graphene from that in other semiconducting materials with parabolic dispersion relations at the band edge.

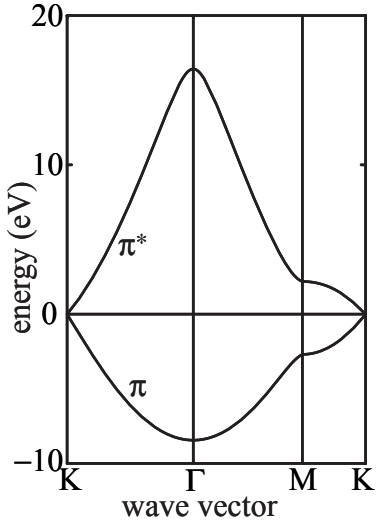


Fig. 3.7 Plot of the π- and the $\pi^\star$-bands in graphene.

3.2 Electron spin and the Zeeman hamiltonian

Spin and magnetic moment. So far we have completely neglected the spin of the electron. However, each electron possesses a magnetic dipole moment μ that can be described by the electronic spin. It is a degree of freedom of the electron, in addition to the three degrees of freedom of the spatial motion. For a free electron, spatial motion and spin dynamics are independent.

In quantum mechanics, the spin operator is defined as

$$\mathbf{S} = \frac{1}{2}\boldsymbol{\sigma},$$

where the components of $\boldsymbol{\sigma}$ are the Pauli matrices

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \qquad (3.14)$$

The magnetic moment of the electron is related to the spin angular momentum via

$$\boldsymbol{\mu} = -\frac{1}{2}g\mu_\mathrm{B}\boldsymbol{\sigma},$$

where $\mu_\mathrm{B} = |e|\hbar/2m_\mathrm{e}$ is Bohr's magneton and $g = 2.0023$. For the electron $\mu_\mathrm{B} = 9.274 \times 10^{-24}$ Am^2 = 57.88 μeV/T.

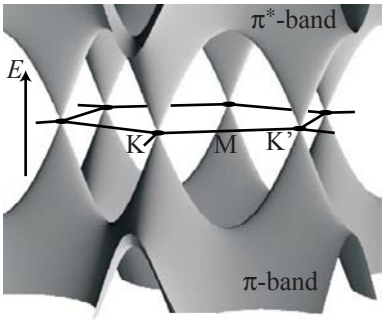


Fig. 3.8 Three-dimensional plot of the π- and the $\pi^\star$-bands in graphene. The black line indicates the boundaries of the hexagonal first Brillouin zone (see also Fig. 3.6).

Spin wave functions are described by two-component spinors. Consistent with the description of the spin operator $\mathbf{S}$ via the Pauli matrices, we write the spinor of an electron in Pauli notation as a two-component vector

$$|\chi\rangle = \begin{pmatrix} \chi_0 \\ \chi_1 \end{pmatrix},$$

where the normalization condition requires $\chi_0^2 + \chi_1^2 = 1$.

Magnetic moments in external magnetic fields. In a homogeneous magnetic field, a torque

$$\mathbf{M} = \boldsymbol{\mu} \times \mathbf{B}$$

acts on the electron and leads to precession about the magnetic field axis. The energy of the magnetic dipole moment in a magnetic field is described by the Zeeman hamiltonian

$$H = -\boldsymbol{\mu}\mathbf{B} = \frac{1}{2} g \mu_{\mathrm{B}} \boldsymbol{\sigma} \mathbf{B}.$$

Example: Spin in a static magnetic field. As an example of the use of Pauli's spinor notation we solve the problem of a spin in a static magnetic field $\mathbf{B} = B(\sin\theta\cos\delta, \sin\theta\sin\delta, \cos\theta)$. The orientation of $\mathbf{B}$ in space is characterized by the two angles θ and δ ($0 \leq \theta \leq \pi$, $0 \leq \delta \leq 2\pi$), which specify a unit vector in the direction of $\mathbf{B}$. The hamiltonian for a spin in this field is given by

$$H = \frac{1}{2} g \mu_{\mathrm{B}} B \begin{pmatrix} \cos\theta & \sin\theta e^{-i\delta} \\ \sin\theta e^{+i\delta} & -\cos\theta \end{pmatrix}.$$

The two energy eigenvalues of this hamiltonian are readily found to be

$$E_{\pm} = \pm \frac{1}{2} g \mu_{\mathrm{B}} B,$$

and the normalized eigenvectors can be written as

$$|\chi_+\rangle = \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2) e^{i\delta} \end{pmatrix}, \quad \text{and} \quad |\chi_-\rangle = \begin{pmatrix} \sin(\theta/2) \\ -\cos(\theta/2) e^{i\delta} \end{pmatrix}.$$

The energy splitting $\Delta E_{\mathrm{Z}} = E_+ - E_- = g\mu_{\mathrm{B}} B$ is called the Zeeman energy. It increases linearly with the magnetic field strength.

Bloch sphere representation. An instructive geometric interpretation of the two eigenvectors is found, if we consider the expectation values of the Pauli matrices. They form the so-called polarization vector $\mathbf{P} = (\langle\sigma_x\rangle, \langle\sigma_y\rangle, \langle\sigma_z\rangle)$. In case of the state $|\chi_+\rangle$ its components are given by

$$\begin{aligned} P_x &= \sin\theta\cos\delta \\ P_y &= \sin\theta\sin\delta \\ P_z &= \cos\theta, \end{aligned}$$

which represents a unit vector in real space which is parallel to **B** (see Fig. 3.9). If a particle occupies this state, we therefore say that its spin is oriented parallel to **B**. For the case of the state $|\chi_-\rangle$ we find

$$\begin{aligned} P_x &= -\sin\theta\cos\delta \\ P_y &= -\sin\theta\sin\delta \\ P_z &= -\cos\theta, \end{aligned}$$

which represents a unit vector antiparallel to **B**, associated with a spin antialigned with the magnetic field (see Fig. 3.9).

The tip of the vector **P** will always lie on the surface of a sphere of unity radius. The representation of a spin state as a point on such a surface as it is shown in Fig. 3.9 is called the *Bloch sphere representation*; the spherical surface itself is called the *Bloch sphere*. In dynamical problems, e.g., in a case where the electron spin is not in an eigenstate of the Zeeman hamiltonian, the time evolution of the spin can be visualized as a trajectory on the Bloch sphere.

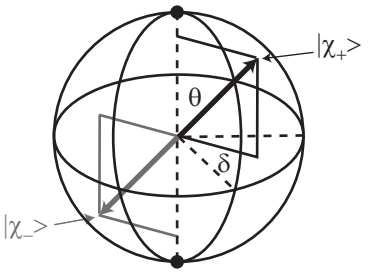


Fig. 3.9 Bloch sphere representation of the electron spin.

The electron spin is a paradigm for a two-level quantum mechanical system. The Bloch-sphere representation of the electron spin can therefore be generalized to a representation of arbitrary two-level quantum systems. In the context of quantum information processing two-level systems are called qubits, for which the Bloch sphere is an illustrative representation. This will be discussed in more detail in section 22.3.2.

3.3 Spin–orbit interaction

Spin in a magnetic field gradient. One variant, how the spin can influence the spatial motion of the electron, occurs in the famous Stern–Gerlach experiment, where the magnetic dipole moment of the electron experiences a force

$$\mathbf{F} = \nabla(\boldsymbol{\mu}\mathbf{B})$$

in a magnetic field gradient. Electrons with different spin orientation are therefore deflected in two opposite directions.

Magnetic moment in external electric fields. Another mutual influence between spin and orbital motion occurs if an electron moves in an electric field. In order to make this more transparent, we use the following plausibility argument: Let us assume that in the inertial system in which an atom (or a crystal lattice) is at rest, there is an electric field **E** caused by the atom (or the lattice). The electron will see in its own rest system not only a pure electric field, but also a magnetic field which is in the lowest order in v/c

$$\mathbf{B}' = -\frac{1}{c^2}\mathbf{v}\times\mathbf{E},$$

due to the relativistic transformation of the fields (primed variables denote quantities in the coordinate system in which the electron is at rest).

The magnetic moments in the lab frame and the electron's rest frame are the same to first order of v/c. The magnetic field $\mathbf{B}'$ couples to the magnetic dipole moment of the electron, i.e., to the spin via the Zeeman interaction and we have in lowest order of v/c

$$\begin{aligned} H'_{SO} &= g\mu_{\mathrm{B}}\mathbf{B}'\mathbf{S}' \\ &= -g\mu_{\mathrm{B}}\frac{1}{c^2}\left(\mathbf{v}\times\mathbf{E}\right)\mathbf{S} = H_{\mathrm{SO}} \end{aligned}$$

The above argument is incomplete as it neglects complications arising due to the acceleration of the electron (see e.g., Jackson, 1983) leading to the so-called Thomas precession (Thomas, 1927). If the term is exactly derived from the relativistic Dirac equation by taking the nonrelativistic limit, the result is

$$H_{\mathrm{SO}} = -\frac{g\mu_{\mathrm{B}}}{2}\frac{1}{c^2}\left(\mathbf{v}\times\mathbf{E}\right)\mathbf{S} \tag{3.15}$$

$$= \frac{g\hbar}{4c^2m_{\mathrm{e}}^2}\left(\nabla V(\mathbf{r})\times\mathbf{p}\right)\mathbf{S}. \tag{3.16}$$

This exact expression differs only by a factor 1/2, the Thomas-factor, from the expression obtained from the above incomplete argument. It is the nature of this spin–orbit interaction that the electron feels a magnetic field oriented normal to its direction of motion and normal to the external electric field. For some purposes it is convenient to combine the expression for the spin–orbit interaction (3.15) with the kinetic energy of the free electron motion giving

$$H = \frac{1}{2m_{\mathrm{e}}}\left[\mathbf{p} - \frac{g\mu_{\mathrm{B}}}{2c^2}(\mathbf{E}\times\mathbf{S})\right]^2,$$

which is correct up to order v/c.

Effect of spin–orbit interaction on the band structure. In general, spin degeneracy of states in a semiconductor is the result of spatial inversion symmetry of the crystal lattice and time-reversal symmetry. Both symmetry operations together transform the wave vector $\mathbf{k}$ into $-\mathbf{k}$. Time reversal, however, also inverts the orientation of the spin. If a crystal lattice possesses a center of inversion, and if time-reversal symmetry is given, the dispersion relations obey $E_\uparrow(\mathbf{k}) = E_\downarrow(\mathbf{k})$, i.e., spin degeneracy is given. This is, for example, the case for the elementary semiconductors silicon and germanium which have a diamond crystal lattice. However, inversion symmetry and time-reversal symmetry do not imply the complete absence of spin–orbit interaction effects.

The strength of the spin–orbit interaction depends on the gradient of the potential and is therefore more important the higher the nuclear charge of the element. Heavy elements in the periodic table show stronger effects. This is also valid in crystals. For example, in silicon the spin–orbit interaction is much weaker than in germanium or gallium arsenide. It is even more important in InAs and InSb.

We have seen above that the periodic pseudopotential lifts degeneracies present in the free electron model at symmetry points in the first Brillouin zone and creates band gaps. The spin–orbit interaction lifts further degeneracies that have remained due to the crystal symmetry. A very important manifestation of this effect is the so-called *spin–orbit split-off band* which is a branch of the valence band lowered energetically due to spin–orbit interaction. For example, in germanium the valence band structure comprises a heavy and a light hole branch degenerate at Γ, and a spin–orbit split-off band that is about $\Delta_0 = 290$ meV lower in energy, while in silicon Δ_0 is only 44 meV. In GaAs we find a bigger value of Δ_0 than in Ge, namely 340 meV. A few values of Δ_0 for selected semiconductors are summarized in Table 3.4.

Table 3.4 Energy difference Δ_0 between the band of heavy and light holes that are degenerate at Γ, and the spin–orbit split-off band for selected semiconductors (Winkler, 2003).

material	Δ_0 (meV)
C	6
Si	44
Ge	290
GaAs	340
InAs	380
GaSb	800
InSb	820
InP	110
AlSb	750
AlAs	290
GaN	11
CdTe	920

If the inversion symmetry of the crystal is broken, such as, for example, in zinc blende semiconductors, i.e., GaAs, InAs, or InSb, the degeneracy $E_\uparrow(\mathbf{k}) = E_\downarrow(\mathbf{k})$ disappears and we talk about the so-called bulk inversion asymmetry (often abbreviated BIA) which adds another contribution to the spin–orbit interaction called the *Dresselhaus contribution* (Dresselhaus, 1955; Winkler, 2003). The dispersion relations have two branches, $E_+(\mathbf{k})$ and $E_-(\mathbf{k})$ (Dresselhaus, 1955). Here, '+' and '−' do not denote the two spin orientations $\uparrow$ and $\downarrow$, because the corresponding states are typically not eigenstates of the spin along a global axis. In time-reversal-invariant systems, i.e., in the absence of a magnetic field, only the more general relation $E_+(\mathbf{k}) = E_-(-\mathbf{k})$ is valid.

3.4 Band structure of some semiconductors

Silicon band structure. The calculated band structure of silicon including the effect of the spin–orbit interaction is depicted in Fig. 3.10(a). Comparing with Fig. 3.3 we find only very small differences owing to the fact that spin–orbit effects are small because silicon is a light element (cf. Table 3.4).

Germanium band structure. The situation is different already for the calculated germanium band structure depicted in Fig. 3.10(b) where the spin–orbit splitting of the valence band states is more than six times bigger than in silicon. Furthermore, in germanium the lowest conduction band minimum occurs at L rather than at X, and an additional conduction band minimum higher in energy arises at Γ.

Gallium arsenide band structure. The calculated band structure of gallium arsenide is depicted in Fig. 3.11(a). The lowest conduction band minimum is at Γ, but higher minima are still present at L and X. The spin–orbit split-off valence band has also moved down in energy as compared to germanium.

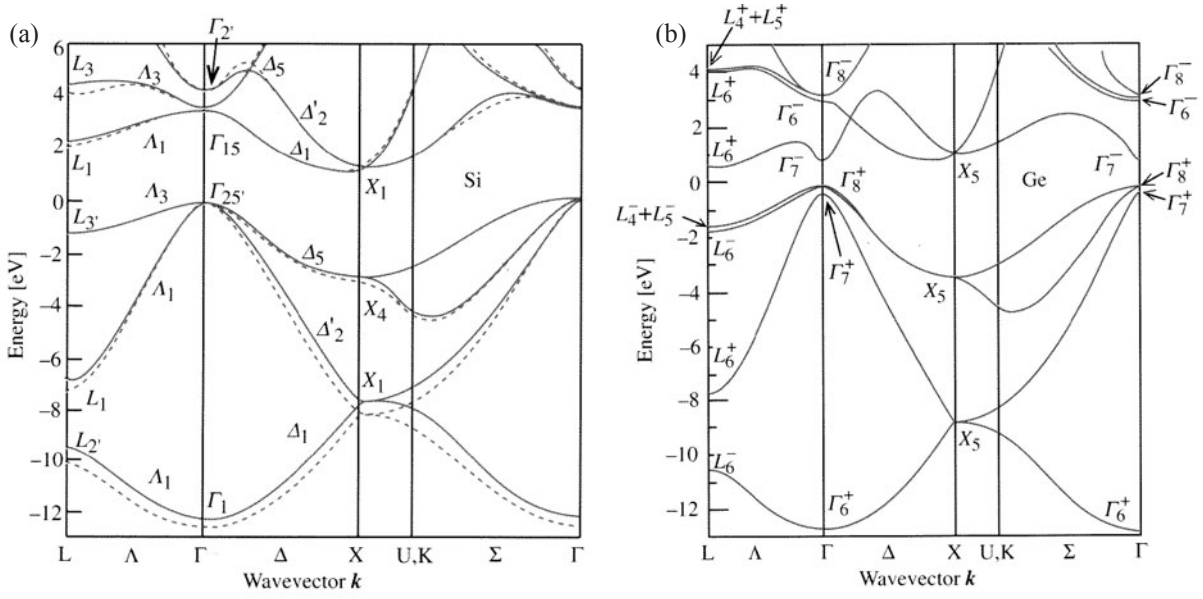


Fig. 3.10 (a) Band structure of silicon resulting from a pseudopotential calculation including the effects of spin–orbit coupling. Solid lines were calculated with a nonlocal, dashed lines with a local pseudopotential. (b) Band structure of germanium obtained as the result of a pseudopotential calculation including spin–orbit coupling effects (Cohen and Chelikowski, 1989).

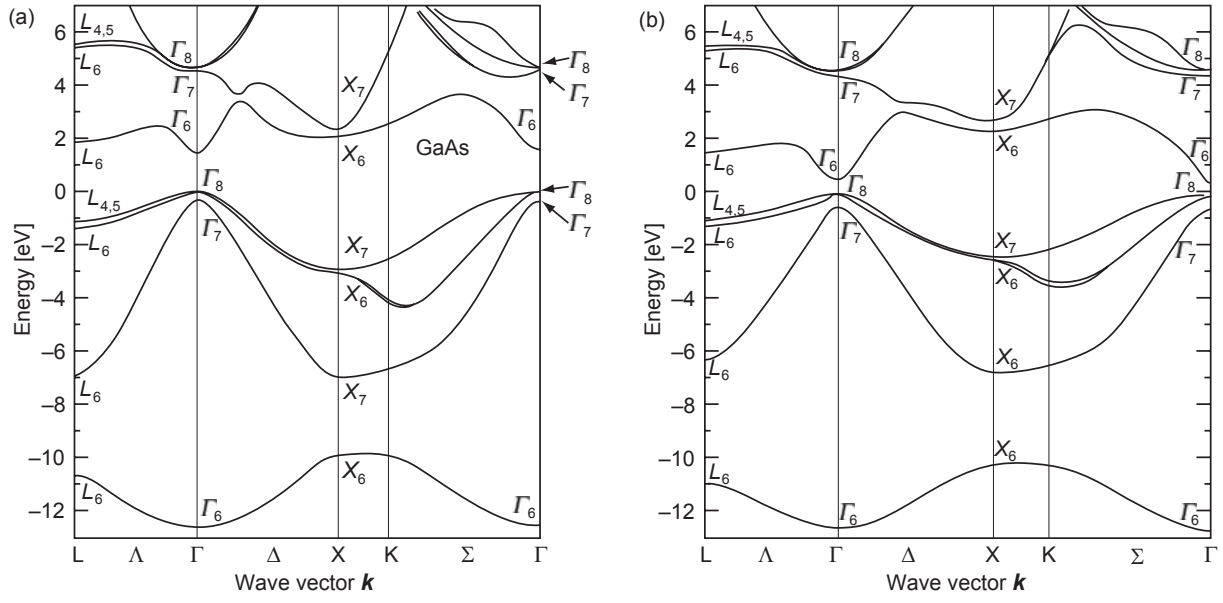


Fig. 3.11 Band structure of GaAs (a) and InAs (b) resulting from calculations using the pseudopotential method including spin–orbit effects (Cohen and Chelikowski, 1989).

Indium arsenide band structure. The indium arsenide band structure depicted in Fig. 3.11(b) is an example where the energy gap ($E_g = 0.4$ eV) is comparable to the spin–orbit splitting of the valence band ($\Delta_0 = 0.38$ eV). As in GaAs, the valence band maximum and the conduction band minimum are at Γ.

Comparison of band structures. All four band structures appear to be very similar to each other and to the band structure of the free electron model in Fig. 3.2. The reasons are the underlying lattice symmetries and there being the same number of valence electrons.

Si and Ge are called *indirect semiconductors*, because the valence band maximum and the conduction band minimum are at different points in the first Brillouin zone. In contrast, GaAs and InAs are called *direct semiconductors*, because the valence band maximum and the conduction band minimum are both at the same point Γ.

3.5 Band structure near band extrema: k·p-theory

In semiconductor nanostructures the relevant parts of the band structure are near the lowest minimum of the conduction band or close to maxima of the valence band. There is a method, called $\mathbf{k}\cdot\mathbf{p}$ perturbation theory, for calculating the band structure close to such extrema. In the following we will give an overview over this widely used and powerful method.

The method. Inserting the wave function of eq. (3.8) into Schrödinger's equation (3.1) for the crystal lattice, we obtain the following equation for the lattice periodic part $u_{n\mathbf{k}}(\mathbf{r})$ of the wave function

$$\left\{\left[\frac{\mathbf{p}^2}{2m_e}+V(\mathbf{r})\right]+\left[\frac{\hbar}{m_e}\mathbf{k}\cdot\mathbf{p}+\frac{\hbar^2k^2}{2m_e}\right]\right\}u_{n\mathbf{k}}(\mathbf{r})=Eu_{n\mathbf{k}}(\mathbf{r}). \qquad (3.17)$$

Here, $\mathbf{p}=-i\hbar\nabla$ is the momentum operator and $u_{n\mathbf{k}}(\mathbf{r})$ fulfills periodic boundary conditions at the boundaries of the primitive cell. Let us assume that we have solved this equation and found the energies E_n and the corresponding functions $u_{n0}(\mathbf{r})\equiv|n\rangle$ for the special case $\mathbf{k}=0$. These functions form a complete set of basis states that can be used to expand $u_{n\mathbf{k}}(\mathbf{r})$ for arbitrary $\mathbf{k}$ leading to

$$u_{n\mathbf{k}}(\mathbf{r})=\sum_n c_n(\mathbf{k})u_{n0}(\mathbf{r}).$$

Inserting this expansion into eq. (3.17) gives the following set of equations determining the coefficients $c_n(\mathbf{k})$:

$$\sum_n\left[\left(E_n+\frac{\hbar^2k^2}{2m_e}\right)\delta_{n,n'}+\frac{\hbar}{m_e}\mathbf{k}\cdot\langle n'|\mathbf{p}|n\rangle\right]c_n(\mathbf{k})=Ec_{n'}(\mathbf{k}) \qquad (3.18)$$

Symmetries of the states at the band edge. Within $\mathbf{k}\cdot\mathbf{p}$-theory the symmetries of the states at the band extrema $u_{n0}(\mathbf{r})$ are of crucial importance. Crystals of semiconductors with diamond structure have the symmetry of the point group O_h of the cube. Zincblende lattices have the symmetry of the point group T_d of the tetrahedron which has a lower symmetry than the cube. We obtain O_h from T_d by adding inversion symmetry. Denoting by C_i the point group containing only the inversion and the identity operation, we obtain $O_h = T_d \otimes C_i$. On the other hand, the point group of the cube is a subgroup of the group $\mathcal{R}$ of arbitrary rotations. This shows that there is a hierarchy of symmetries that can be expressed as

$$\mathcal{R} \supset O_h \supset T_d.$$

The hamiltonian H of the crystal can be split into parts with these hierarchical symmetries, as

$$H = H_{\text{rotation}} + H_{\text{cube}} + H_{\text{tetrahedron}},$$

where H_{rotation} is the spherically symmetric part of the hamiltonian, H_{cube} is the part with cubic symmetry, and $H_{\text{tetrahedron}}$ is the part with the symmetry of the tetrahedron. Splitting the hamiltonian in this fashion provides a hierarchy of approximations. In the *spherical approximation* one keeps only the part H_{rotation}, and the eigenstates of the hamiltonian are eigenstates of total angular momentum $\mathbf{J}^2$ and its z-component J_z. Neglecting the spin, this leads to eigenstates of orbital angular momentum and we can talk about s-like or p-like states at the band edge.

In diamond and zincblende semiconductors the states at the valence band edge have p-like symmetry. As a consequence there are three degenerate angular momentum states $|\,\ell = 1, \ell_z = 0, \pm 1\rangle$. These (orbital) states are frequently denoted as

$$|\,X\rangle = \frac{-1}{\sqrt{2}}\left(|\,1,1\rangle - |\,1,-1\rangle\right), |\,Y\rangle = \frac{i}{\sqrt{2}}\left(|\,1,1\rangle + |\,1,-1\rangle\right), |\,Z\rangle = |\,1,0\rangle\,.$$

The phase for the orbital states is chosen such that these are real-valued functions.

In contrast, the state at the conduction band edge at Γ has s-like symmetry ($\ell = 0$) and is denoted as $|\,S\rangle$. We choose this function to be purely imaginary. The next higher conduction band is again p-like with the three states

$$|\,X'\rangle\,, |\,Y'\rangle\,, |\,Z'\rangle\,,$$

which are again chosen to be purely imaginary. All these states can be occupied with two spin orientations.

The spherical approximation describes quite well the relevant big energy scales, such as the energetic spacing of states at Γ. Taking H_{cube} into account leads to the angular momentum being only an 'almost good' quantum number. This means that angular momentum eigenstates will

mix and the energies will be slightly modified, but the changes are small on the scale of the interband separation. Adding $H_{\text{tetrahedron}}$ causes even smaller corrections. This hierarchy manifests itself in the close similarities between the band structures of diamond lattices (Si, Ge) and zincblende lattices.

Band edge parameters. In $\mathbf{k}\cdot\mathbf{p}$-theory the eigenvalue problem is expressed using matrix elements of the momentum operator. Due to the symmetry of the wave functions we find

$$\begin{aligned}
\frac{m_\mathrm{e}}{\hbar}P &= \langle X\,|p_x|\,S\rangle = \langle Y\,|p_y|\,S\rangle = \langle Z\,|p_z|\,S\rangle\,,\\
\frac{m_\mathrm{e}}{\hbar}P' &= \langle X'\,|p_x|\,S\rangle = \langle Y'\,|p_y|\,S\rangle = \langle Z'\,|p_z|\,S\rangle\,,\\
\frac{m_\mathrm{e}}{\hbar}Q &= \langle X\,|p_y|\,Z'\rangle = \langle Y\,|p_z|\,X'\rangle = \langle Z\,|p_x|\,Y'\rangle\,,
\end{aligned}$$

with the so-called band edge parameters P, P' and Q. Other matrix elements, such as $\langle X\,|p_y|\,S\rangle$ and others, are zero. In our notation P and Q are real, while P' is purely imaginary.

The only additional parameters of the theory are the band edge energies E_n. The upper edge of the valence band is chosen to be the zero of energy by convention; the lower edge of the conduction band (at Γ) has the energy E_0 and the higher conduction band (at Γ) has energy E_0'.

Perturbation theory. For small $|\mathbf{k}|$ in eq. (3.18) we can treat the $\mathbf{k}$-dependent terms as a perturbation and calculate the energy dispersions using perturbation theory. This method is called $\mathbf{k}\cdot\mathbf{p}$-perturbation theory. There will be no terms linear in $\mathbf{k}$ at band extrema, i.e., the corrections to E_n vanish in first order. In second order, we obtain, for nondegenerate E_n, the expression

$$E_n(\mathbf{k}) = E_n + \frac{\hbar^2k^2}{2m_\mathrm{e}} + \frac{\hbar^2}{m_\mathrm{e}^2}\sum_{m,m\neq n}\frac{|\mathbf{k}\cdot\mathbf{p}_{mn}|^2}{E_n - E_m}.$$

Typically many of the matrix elements $\mathbf{p}_{mn} = 0$, and the last term of this equation simplifies considerably.

Following this treatment, the wave functions $u_{n\mathbf{k}}$ close to a band extremum are given in first order as

$$u_{n\mathbf{k}}(\mathbf{r}) = u_{n0}(\mathbf{r}) + \frac{\hbar}{m_\mathrm{e}}\sum_{n'\neq n}\frac{\mathbf{k}\cdot\mathbf{p}_{nn'}}{E_n - E_{n'}}\cdot u_{n'0}(\mathbf{r}). \tag{3.19}$$

Conduction band dispersion of GaAs. As an example, we consider the conduction band minimum of GaAs with the energy parameters $E_0 = E_\mathrm{g} = 1.519$ eV and $E_0' = 4.488$ eV. The largest contributions to the last term arise from the energetically closest bands, i.e., from the valence band. Taking $\mathbf{k}$ in x-direction, only the matrix element

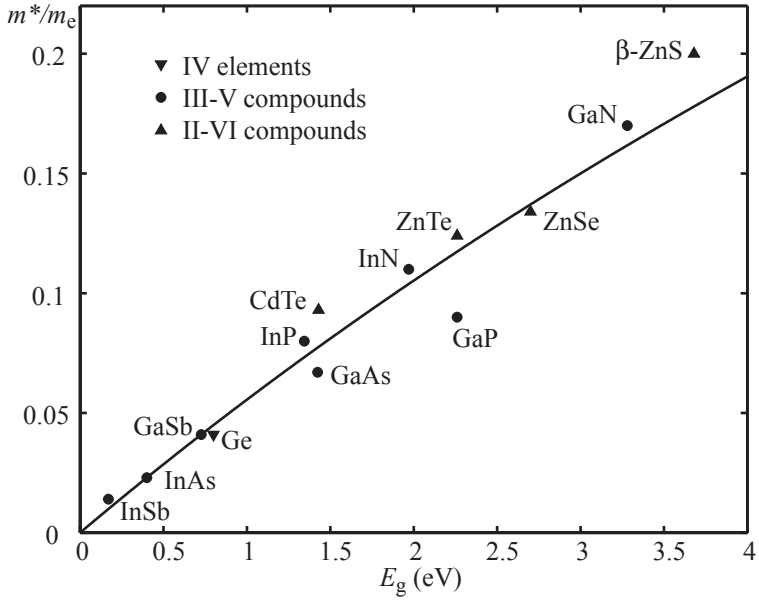


Fig. 3.12 Band gap E_g vs. relative effective conduction band mass $m^\star/m_e$ for a number of semiconductors. The solid line represents the result of eq. (3.21) with $2m_eP^2/\hbar^2 = 17\,\mathrm{eV}$.

$< S|p_x|X > \equiv m_eP/\hbar$ is nonzero and we get the approximate dispersion

$$E_c(\mathbf{k}) \approx E_c + \frac{\hbar^2k^2}{2m_e} + \frac{\hbar^2}{m_e^2}\frac{|k(m_e/\hbar)P|^2}{E_c - E_v} = E_c + \frac{\hbar^2k^2}{2m_e}\left(1 + \frac{2m_eP^2/\hbar^2}{E_g}\right). \tag{3.20}$$

Conduction band effective mass. The dispersion remains parabolic like the dispersion of a free electron. However, the curvature of the parabola, described by the electron's mass in the free electron case, is modified. It is therefore convenient to introduce the conduction band *effective mass* parameter $m^\star$ as

$$\frac{1}{m^\star} = \frac{1}{m_e}\left(1 + \frac{2m_eP^2/\hbar^2}{E_g}\right), \tag{3.21}$$

leading to the dispersion

$$E_c(\mathbf{k}) \approx E_c + \hbar^2k^2/2m^\star. \tag{3.22}$$

We see from eq. (3.21) that the effective mass in semiconductors with large band gap E_g tends to be bigger than in those with small band gaps. Figure 3.12 shows the relation between the band gap and the effective conduction band masses at Γ of a number of semiconductors (symbols), together with the approximation given in eq. (3.21). In fact, it turns out that the momentum parameter P is very similar for different materials. We can estimate its size by considering how the band structure of GaAs in Fig. 3.11(a) comes about. The parameter P is essentially the expectation value of the momentum in the vicinity of Γ. The free electron

dispersion has been folded back once at the boundary of the first Brillouin zone leading to a wave vector $2\pi/a$ at Γ. With $a \approx 0.5$ nm this gives $2m_e P^2/\hbar^2 \approx 22$ eV. With this estimate of P we obtain a value $m^\star = 0.061 m_e$ which is very close to the measured $m^\star = 0.067 m_e$. The effective conduction band masses for a number of semiconductors are listed in Table 3.5.

Table 3.5 Effective conduction band masses of some semiconductors (in units of the free electron mass).

material	$m^\star/m_e$
GaN	0.17
GaAs	0.067
GaSb	0.047
InP	0.080
InAs	0.023
InSb	0.014

Constant energy surface for isotropic dispersions. In general, surfaces of constant energy in conduction and valence bands of semiconductors play an important role for many physical phenomena. For example, if a semiconductor is strongly doped, a Fermi surface arises near the valence band maximum or conduction band minimum—very much like in a metal. It plays an important role in the conductivity at low temperatures. In the case of an isotropic dispersion like eq. (3.20), surfaces of constant energy are simply spherical. The corresponding Fermi surface is referred to as a 'Fermi sphere'.

Density of states for parabolic dispersions in three dimensions. As with the shape of constant energy surfaces, the density of states is often an important quantity entering certain physical properties. The general definition of the density of states is

$$\mathcal{D}(E) = \frac{1}{V} \sum_{n,\mathbf{k},\sigma} \delta(E - E_{n\mathbf{k}\sigma}),$$

where V is the volume of the crystal, n is the band index, $\mathbf{k}$ is the wave vector and σ is the spin quantum number. The quantity $\mathcal{D}(E)dE$ describes the number of quantum states in the energy interval $[E, E+dE]$ normalized to the volume. The functional form of the density of states depends only on the dispersion relation $E_{n\mathbf{k}\sigma}$. Here we assume that states are spin degenerate. From its definition we see that the density of states is a sum of contributions of the individual energy bands:

$$\mathcal{D}(E) = 2 \sum_n \mathcal{D}_n(E), \text{ where } \mathcal{D}_n(E) = \frac{1}{V} \sum_{\mathbf{k}} \delta(E - E_{n\mathbf{k}}).$$

The prefactor 2 is the result of the spin degeneracy.

Integrating the density of states over energy up to a maximum energy value results in the total number of states (per volume) below this energy:

$$\int_{-\infty}^{E} dE' \mathcal{D}(E') = \frac{2}{V} \int_{-\infty}^{E} dE' \sum_{n,\mathbf{k}} \delta(E' - E_{n\mathbf{k}}) = \frac{2}{V} \sum_{\substack{n,\mathbf{k} \\ E_{n\mathbf{k}} < E}} 1 = \mathcal{N}(E).$$

Correspondingly,

$$\mathcal{D}(E) = \frac{d\mathcal{N}(E)}{dE}.$$

This relation is frequently used for calculating the density of states.

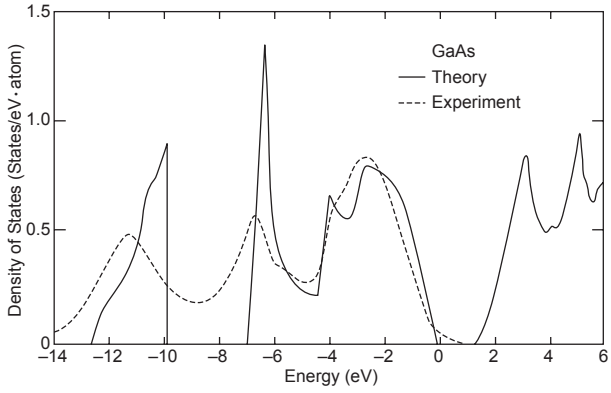


Fig. 3.13 Density of states of GaAs. The dotted line is the result of an XPS-measurement, and the solid line is the prediction of a pseudopotential calculation (Cohen and Chelikowski, 1989).

For complicated functional forms of the dispersion relation the density of states can typically not be calculated analytically. As an example, Fig. 3.13 shows the density of states of GaAs over a broad energy range. The band gap can be clearly identified as an energy interval with vanishing density of states.

Here we are interested in the density of states for the case of isotropic parabolic dispersion, eq. (3.22), giving

$$\begin{aligned}\mathcal{N}_c(E) &= \frac{2}{V}\sum_{\mathbf{k},E_{n\mathbf{k}}<E} 1 = \frac{2}{(2\pi)^3}\int_0^{k(E)} d^3k \\ &= \frac{8\pi}{(2\pi)^3}\int_0^{k(E)} dk k^2 = \frac{8\pi}{(2\pi)^3}\int_0^{k(E)} d(k^2)\cdot\frac{k}{2}\end{aligned}$$

Using the dispersion relation, the magnitude of the wave vector $\mathbf{k}$ can be expressed as a function of energy as $k^2 = 2m^\star(E - E_\mathrm{c})/\hbar^2$. This leads to

$$\mathcal{N}_\mathrm{c}(E) = \frac{2}{(2\pi)^2}\left(\frac{2m^\star}{\hbar^2}\right)^{3/2}\int_0^{E-E_\mathrm{c}} dE\cdot\sqrt{E}.$$

and the density of states is

$$\mathcal{D}_\mathrm{3D}(E) = \frac{d\mathcal{N}_\mathrm{c}(E)}{dE} = \frac{2}{(2\pi)^2}\left(\frac{2m^\star}{\hbar^2}\right)^{3/2}\sqrt{E-E_\mathrm{c}}. \tag{3.23}$$

Extensions of simple $\mathbf{k}\cdot\mathbf{p}$-theory. The $\mathbf{k}\cdot\mathbf{p}$-perturbation theory can be extended to the case of degenerate band edge states E_n, e.g., for calculating the dispersion relation at the valence band edge. However, in this case spin–orbit effects are important (see below). The method can also be extended to higher orders resulting in nonparabolicities of bands. Furthermore it can be performed also at conduction band minima that do not arise at $\mathbf{k} = 0$, as found, for example, in silicon and germanium.

Conduction band dispersion for silicon. The dispersion of the conduction band in silicon is particularly interesting near the six equivalent

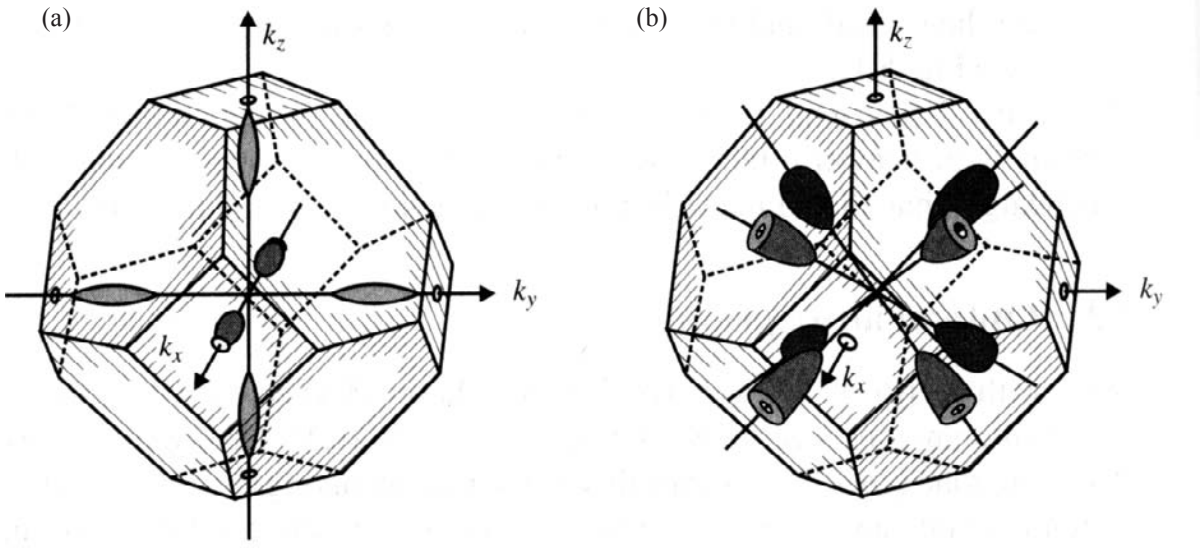


Fig. 3.14 (a) Surfaces of constant energy close to the minima of the conduction band of silicon. (b) Surfaces of constant energy close to the conduction band of germanium (Singleton, 2001).

X-points at the edge of the first Brillouin zone (cf. Fig. 3.1). It is found that the dispersion is not isotropic around the conduction band minima, but follows the dispersion relation (here quoted for the minimum in the k_x-direction)

$$E_\mathrm{c} \approx \frac{\hbar^2(k_x - k_0)^2}{2m_\mathrm{L}} + \frac{\hbar^2(k_y^2 + k_z^2)}{2m_\mathrm{T}}, \tag{3.24}$$

with $m_\mathrm{L} = 0.98m_\mathrm{e}$, $m_\mathrm{T} = 0.19m_\mathrm{e}$, and k_0 being the position of the minimum in the k_x direction.

Constant energy surfaces in the silicon conduction band. Following the above dispersion relation, we can see that surfaces of constant energy close to the conduction band minima of silicon are ellipsoids. Their arrangement in reciprocal space reflects the symmetry of the crystal [see Fig. 3.14(a)].

Conduction band dispersion for germanium. In germanium, conduction band minima arise at the eight equivalent L points of the first Brillouin zone (cf. Fig. 3.1). Around these minima, the dispersion is anisotropic, as in the case of silicon. A longitudinal effective mass $m_\mathrm{L} = 1.64m_\mathrm{e}$ and a transverse effective mass $m_\mathrm{T} = 0.082m_\mathrm{e}$ are found. The dispersion is described by eq. (3.24), but the orientation of $\mathbf{k}_0$ has to be changed to the (1,1,1) or equivalent directions.

Constant energy surfaces in the germanium conduction band. The eight equivalent surfaces of constant energy in the germanium conduction band are of ellipsoidal shape. They are cut in the middle by the Brillouin zone edge as shown in Fig. 3.14(b).

Dispersion of graphene near K. For the calculation of the dispersion of graphene near $\mathbf{K}$ [see also (Ando, 2005)] we start from eq. (3.17) and assume that we have solved this equation for $\mathbf{k} = \mathbf{K}$ and found the lattice periodic functions $u_{n\mathbf{K}}(\mathbf{r})$. Within the tight-binding approximation introduced for graphene on page 23 we can write the two degenerate wave functions at $\mathbf{K}$ as

$$\begin{aligned} u_{\mathbf{K}}^{(\mathrm{A})}(\mathbf{r}) &= \frac{1}{\sqrt{N}} \sum_{\mathbf{R}} e^{-i\mathbf{K}(\mathbf{r}-\mathbf{R})} \phi(\mathbf{r}-\mathbf{R}-\mathbf{t}_1) \\ u_{\mathbf{K}}^{(\mathrm{B})}(\mathbf{r}) &= \frac{1}{\sqrt{N}} \sum_{\mathbf{R}} e^{-i\mathbf{K}(\mathbf{r}-\mathbf{R})} \phi(\mathbf{r}-\mathbf{R}), \end{aligned} \tag{3.25}$$

where the former is nonzero only on sites of A-atoms, the latter only on sites of B-atoms.

Now we consider the problem at $\mathbf{k} = \mathbf{K} + \mathbf{q}$ for small $\mathbf{q}$. Inserting into eq. (3.17) the eigenvalue equation for the $u_{n\mathbf{q}}(\mathbf{r})$ gives

$$\left\{ \left[\frac{\mathbf{p}^2}{2m_\mathrm{e}} + V(\mathbf{r}) \right] + \left[\frac{\hbar}{m_\mathrm{e}} \mathbf{K}\mathbf{p} + \frac{\hbar^2 \mathbf{K}^2}{2m_\mathrm{e}} \right] + \frac{\hbar}{m_\mathrm{e}} \mathbf{q}\,(\mathbf{p} + \hbar\mathbf{K}) + \frac{\hbar^2 \mathbf{q}^2}{2m_\mathrm{e}} \right\} u_{n\mathbf{q}} = E u_{n\mathbf{q}}.$$

Because we are interested only in small $\mathbf{q}$, we neglect the $\mathbf{q}^2$-term on the left-hand side. The solutions are expanded in the eigenfunctions (3.25) according to

$$u_{n\mathbf{q}}(\mathbf{r}) = A_{\mathbf{q}} u_{\mathbf{K}}^{(A)}(\mathbf{r}) + B_{\mathbf{q}} u_{\mathbf{K}}^{(B)}(\mathbf{r}).$$

Contributions of other bands are neglected in lowest order.

Inserting this wave function into the eigenvalue equation and projecting onto the two basis functions $u_{\mathbf{K}}^{(A)}$ and $u_{\mathbf{K}}^{(B)}$ gives the matrix equation for the two coefficients $A_{\mathbf{q}}$ and $B_{\mathbf{q}}$

$$\frac{\hbar}{m_\mathrm{e}} \mathbf{q} \begin{pmatrix} \mathbf{p}_{AA} + \hbar\mathbf{K} & \mathbf{p}_{AB} \\ \mathbf{p}_{BA} & \mathbf{p}_{BB} + \hbar\mathbf{K} \end{pmatrix} \begin{pmatrix} A_{\mathbf{q}} \\ B_{\mathbf{q}} \end{pmatrix} = E \begin{pmatrix} A_{\mathbf{q}} \\ B_{\mathbf{q}} \end{pmatrix},$$

where $\mathbf{p}_{AA}$, $\mathbf{p}_{BB}$, and $\mathbf{p}_{AB}$ are matrix elements of the momentum operator. Neglecting overlap integrals between neighboring sites, we find $\mathbf{p}_{AA} = \mathbf{p}_{BB} = -\hbar\mathbf{K}$. For the evaluation of $\mathbf{p}_{AB}$ we take only nearest neighbor contributions into account and assume circular symmetry of $\phi(\mathbf{r})$. This leads to (cf. Fig. 3.5 for the definitions of the vectors $\mathbf{t}_\ell$)

$$\mathbf{p}_{AB} = \frac{\hbar}{i} \lambda \sum_{\ell=1}^{3} e^{i\mathbf{K}(\mathbf{t}_1 - \mathbf{t}_\ell)} \mathbf{t}_\ell = \frac{\sqrt{3}\hbar\lambda}{2} (\mathbf{e}_x - i\mathbf{e}_y),$$

where λ is a coupling constant. The eigenvalue problem at $\mathbf{K}$ therefore reduces to

$$\frac{\hbar^2}{2m_\mathrm{e}}\sqrt{3}\lambda \begin{pmatrix} 0 & q_x - iq_y \\ q_x + iq_y & 0 \end{pmatrix} \begin{pmatrix} A_\mathbf{q} \\ B_\mathbf{q} \end{pmatrix} = E \begin{pmatrix} A_\mathbf{q} \\ B_\mathbf{q} \end{pmatrix}. \qquad (3.26)$$

The solution of this eigenvalue problem gives the energy dispersion

$$E(\mathbf{q}) = \pm\hbar c^\star |\mathbf{q}|,$$

where we have introduced the effective velocity $c^\star = \hbar\sqrt{3}\lambda/2m_\mathrm{e}$ with the value $c^\star \approx 10^6\,\mathrm{m/s}$. This velocity can be regarded as the only band structure parameter which is relevant in grapene near $\mathbf{K}$ for the band under consideration. In the expression for the dispersion, the '+'-sign refers to the conduction band ($\pi^\star$-band) dispersion, whereas the '−'-sign refers to the valence band (π-band, see also Fig. 3.7).

Using the vector σ of Pauli matrices, the effective hamiltonian in the vicinity of $\mathbf{K}$ [see eq. (3.26)] can be written as $H_\mathbf{K} = \hbar c^\star \mathbf{q}\boldsymbol{\sigma} = \hbar c^\star |\mathbf{q}| \mathbf{n}\boldsymbol{\sigma}$ with $\mathbf{n} = \mathbf{q}/|\mathbf{q}|$. Owing to the two-component state vector $(A_\mathbf{q}, B_\mathbf{q})$, the electrons near $\mathbf{K}$ are often said to have a pseudospin which gives the relative amplitudes of the electronic wave function on the two sublattice atoms. The direction of the pseudospin determines the character of the underlying molecular orbital state, e.g., bonding or antibonding. The hamiltonian $H_\mathbf{K}$ shows that the direction of the pseudospin is always tied to the direction of $\mathbf{q}$, by analogy with the physical spin of a massless neutrino which points along the direction of propagation. The operator $\mathbf{n}\boldsymbol{\sigma}$ is the operator of the helicity of a particle with zero rest mass. It has eigenvalues ± 1, called right-handed ('+') and left-handed ('−') helicity. At $\mathbf{K}$, conduction band states are right-handed (positive helicity), whereas valence band states are left-handed (negative helicity).

Density of states in graphene. The density of states in graphene around the energy of the $\mathbf{K}$-point differs from that of three-dimensional, and two-dimensional systems with parabolic dispersion relations. The number of states per unit area in the conduction band below the energy E is given by

$$\mathcal{N}_\mathrm{c}(E) = \frac{4}{A} \sum_{\mathbf{k}, E_\mathbf{k} < E} 1 = \frac{4}{(2\pi)^2} \int_0^{k(E)} d^2k.$$

The prefactor 4 results from the two-fold spin, and the two-fold valley degeneracy ($\mathbf{K}$ and $\mathbf{K}'$). The integral evaluates to

$$\mathcal{N}_\mathrm{c}(E) = \frac{4}{2\pi} \int_0^{k(E)} dkk = \frac{4}{2\pi\hbar^2 c^{\star 2}} \int_0^E dEE.$$

As a consequence, the density of states in the conduction band ($E > 0$) is

$$\mathcal{D}_\mathrm{c}(E) = \frac{2|E|}{\pi\hbar^2 c^{\star 2}}. \qquad (3.27)$$

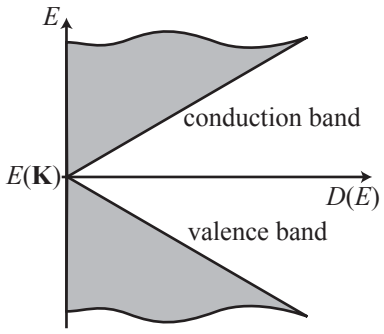


Fig. 3.15 Density of states of graphene near the energy of the **K**-point.

The valence band is symmetric to the conduction band and gives the density of states for $E < 0$

$$\mathcal{D}_{\mathrm{v}}(E) = \frac{2|E|}{\pi\hbar^2 c^{\star 2}}.$$

The total density of states of graphene is depicted in Fig. 3.15. We see that the total density of states at the energy of the **K**-point vanishes for ideal graphene.

3.6 Spin–orbit interaction within k·p-theory

Adding spin–orbit terms to the $\mathbf{k}\cdot\mathbf{p}$-method. Spin–orbit interaction is introduced into $\mathbf{k}\cdot\mathbf{p}$-theory by adding the spin–orbit coupling hamiltonian from eq. (3.16) to the hamiltonian in eq. (3.1). The wave functions of the system are now spinors with two components with a combined band–spin index ν [cf. eq. (3.8)]:

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\mathbf{r}} u_{\nu\mathbf{k}}(\mathbf{r})$$

Inserting this expression into Schrödinger's equation gives

$$\left\{\left[\frac{\mathbf{p}^2}{2m_{\mathrm{e}}} + V(\mathbf{r}) + \frac{g\hbar}{4c^2 m_{\mathrm{e}}^2}\left(\nabla V(\mathbf{r})\times\mathbf{p}\right)\mathbf{S}\right] + \left[\frac{\hbar}{m_{\mathrm{e}}}\mathbf{k}\cdot\boldsymbol{\pi} + \frac{\hbar^2 k^2}{2m_{\mathrm{e}}}\right]\right\} u_{\nu\mathbf{k}}(\mathbf{r}) = E u_{\nu\mathbf{k}}(\mathbf{r}), \quad (3.28)$$

with

$$\boldsymbol{\pi} = \mathbf{p} + \frac{g\hbar}{4m_{\mathrm{e}}c^2}\mathbf{S}\times\nabla V(\mathbf{r}).$$

Again, we first consider the problem at $\mathbf{k} = 0$ and obtain the dispersion relations for small k in a second step from perturbation theory.

Band edge states and their symmetries. Owing to the spin–orbit coupling term, the determination of the spinors $u_{\nu 0}(\mathbf{r})$ from

$$\left[\frac{\mathbf{p}^2}{2m_{\mathrm{e}}} + V(\mathbf{r}) + \frac{g\hbar}{4c^2 m_{\mathrm{e}}^2}\left(\nabla V(\mathbf{r})\times\mathbf{p}\right)\mathbf{S}\right] u_{\nu 0}(\mathbf{r}) = E u_{\nu 0}(\mathbf{r})$$

is not straightforward. We choose the basis functions $u_{n0}(\mathbf{r}) \otimes |\,\sigma\rangle$ of the problem without spin–orbit coupling, where $S_z\,|\,\sigma\rangle = \sigma\,|\,\sigma\rangle$ with $\sigma = \pm 1/2$. Some freedom remains in the combination of the degenerate band edge states, e.g., $|\,X\rangle$, $|\,Y\rangle$, and $|\,Z\rangle$. For the nondegenerate lowest conduction band we choose the two basis states

$$\left|\begin{array}{c} S \\ 0 \end{array}\right\rangle \text{ and } \left|\begin{array}{c} 0 \\ S \end{array}\right\rangle.$$

Following the procedure common in atom physics, the p-like states can first be combined to states with well-defined angular momentum $|\,\ell, \ell_z\rangle$.

This results in the three $\ell = 1$-like states $|1,1\rangle = -(|X\rangle + i\,|Y\rangle)/\sqrt{2}$, $|1,0\rangle = |Z\rangle$, and $|1,-1\rangle = (|X\rangle - i\,|Y\rangle)/\sqrt{2}$. Again following procedures known from atom physics, we further define eigenstates of total angular momentum $\mathbf{J} = \mathbf{L} + \mathbf{S}$ resulting in six states, four of which are $j = 3/2$-like, namely,

$$\begin{aligned}
\left|\frac{3}{2},\frac{3}{2}\right\rangle &= -\frac{1}{\sqrt{2}}\left|\begin{array}{c} X+iY \\ 0 \end{array}\right\rangle \\
\left|\frac{3}{2},\frac{1}{2}\right\rangle &= \frac{1}{\sqrt{6}}\left|\begin{array}{c} 2Z \\ -X-iY \end{array}\right\rangle \\
\left|\frac{3}{2},-\frac{1}{2}\right\rangle &= \frac{1}{\sqrt{6}}\left|\begin{array}{c} X-iY \\ 2Z \end{array}\right\rangle \\
\left|\frac{3}{2},-\frac{3}{2}\right\rangle &= \frac{1}{\sqrt{2}}\left|\begin{array}{c} 0 \\ X-iY \end{array}\right\rangle
\end{aligned}$$

and two of which are $j = 1/2$-like, i.e.,

$$\begin{aligned}
\left|\frac{1}{2},\frac{1}{2}\right\rangle &= -\frac{1}{\sqrt{3}}\left|\begin{array}{c} Z \\ X+iY \end{array}\right\rangle \\
\left|\frac{1}{2},-\frac{1}{2}\right\rangle &= -\frac{1}{\sqrt{3}}\left|\begin{array}{c} X-iY \\ -Z \end{array}\right\rangle .
\end{aligned}$$

Corresponding definitions are made for the band edge states of the p-like upper conduction band.

Matrix equation for determining dispersion relations. Having defined these basis states, we can expand the lattice periodic spinors $u_{\nu\mathbf{k}}$:

$$u_{\nu\mathbf{k}}(\mathbf{r}) = \sum_{n,\sigma} c_{n,\sigma}(\mathbf{k}) u_{n0} \otimes |\,\sigma\rangle = \sum_{n,\sigma} c_{n,\sigma}(\mathbf{k})\,|\,n,\sigma\rangle\ .$$

With this expansion we now substitute in eq. (3.28) and obtain the matrix equation for the expansion coefficients $c_{n,\sigma}(\mathbf{k})$:

$$\begin{aligned}
\sum_{n',\sigma'} \Bigg\{ \left[E_{n'} + \frac{\hbar^2 k^2}{2m_\mathrm{e}}\right] \delta_{n,n'}\delta_{\sigma,\sigma'} + \Delta_{n\sigma,n'\sigma'} \\
+ \frac{\hbar}{m_\mathrm{e}} \mathbf{k}\cdot\mathbf{P}_{n\sigma,n'\sigma'} \Bigg\} c_{n',\sigma'}(\mathbf{k}) &= E_n(\mathbf{k}) c_{n,\sigma}(\mathbf{k}).
\end{aligned}$$

Here we have introduced

$$\begin{aligned}
\Delta_{n\sigma,n'\sigma'} &= \frac{g\hbar}{4c^2 m_\mathrm{e}^2} \langle n,\sigma\,|(\nabla V(\mathbf{r}) \times \mathbf{p})\,\mathbf{S}|\,n'\sigma'\rangle \\
\mathbf{P}_{n\sigma,n'\sigma'} &= \langle n,\sigma\,|\boldsymbol{\pi}|\,n',\sigma'\rangle\,.
\end{aligned}$$

Band edge parameters. The spin–orbit-related matrix element $\Delta_{n'\sigma',n\sigma}$ leads to coupling and splitting of band edge states at $k = 0$.

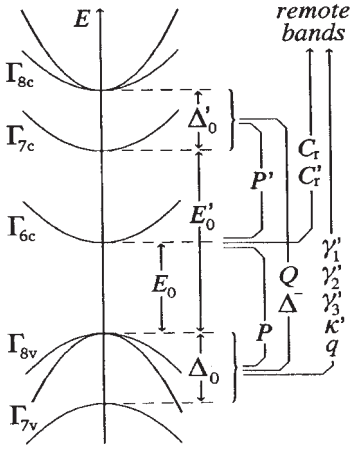


Fig. 3.16 Graphical overview of band edge parameters. (Reprinted with permission from Mayer and Roessler, 1991. Copyright 1991 by the American Physical Society.)

However, owing to the symmetry of states, many of these matrix elements are zero. For example, the band edge state of the GaAs valence band is six-fold degenerate (three orbital wave functions with two spin states each). The spin–orbit interaction splits these six states at $k = 0$ into four plus two, where the latter two form the spin–orbit split-off band. The corresponding band edge parameter is the spin–orbit gap (cf. Table 3.4)

$$\Delta_0 = -\frac{3i\hbar}{4m_e^2c^2}\left\langle X\left|[(\nabla V)\times\mathbf{p}]_y\right|Z\right\rangle.$$

For the two-fold spin degenerate s-like conduction band all matrix elements $\Delta_{c\sigma,c\sigma'} = 0$. For the p-like higher conduction band there is a spin–orbit gap Δ_0' as for the valence band with

$$\Delta_0' = -\frac{3i\hbar}{4m_e^2c^2}\left\langle X'\left|[(\nabla V)\times\mathbf{p}]_y\right|Z'\right\rangle.$$

The nonvanishing off-diagonal matrix element $\Delta_{n'\sigma',n\sigma}$ couples at $\mathbf{k} = 0$ the p-like valence band states to the p-like higher conduction band states via the matrix element

$$\Delta^- = -\frac{3i\hbar}{4m_e^2c^2}\left\langle X\left|[(\nabla V)\times\mathbf{p}]_y\right|Z'\right\rangle,$$

if the crystal has no inversion symmetry, i.e., in zincblende crystals. Within our definition, the matrix elements P, Q, Δ_0, and Δ_0' are real, whereas P' and Δ'_- are purely imaginary.

The matrix elements $\mathbf{P}_{n\sigma,n'\sigma'}$ mix band edge states more strongly the larger $\mathbf{k}$ is and the smaller $|E_n - E_{n'}|$ is. Without spin–orbit coupling this leads to the renormalization of the electron mass, i.e., to the concept of the effective mass. In addition, spin–orbit interaction leads to an increased mixing between the two spin states with increasing k.

Figure 3.16 shows an overview of the various band edge parameters and their influence on band structure. These parameters are considered in the framework of the so-called 'extended Kane model', which is based on the 14×14 matrix hamiltonian which results from consideration of the basis states at the valence and conduction band extrema introduced above. Table 3.6 lists values of the band edge parameters for selected semiconductors.

Coupling of the s-like conduction band to remote bands is treated here via the reduced Hermann–Weisbuch parameters C_r und C_r' in perturbation theory. The corresponding coupling of the valence band states to remote bands is similarly treated using the reduced Luttinger parameters γ_i', κ' and q'.

Spin–orbit interaction hamiltonian in zincblende crystals. Using the above techniques it can be shown that the spin–orbit interaction can be incorporated in the description of conduction band electrons in zincblende crystals by considering the additional hamiltonian (D'yakonov and Perel, 1972; Winkler, 2003)

$$H_D \propto p_x(p_y^2 - p_z^2)\sigma_x + p_y(p_z^2 - p_x^2)\sigma_y + p_z(p_x^2 - p_y^2)\sigma_z. \qquad (3.29)$$

Table 3.6 Band edge parameters for selected semiconductors (Winkler, 2003).

Material	GaAs	AlAs	InAs	InSb	InP
$m^\star/m_\mathrm{e}$	0.067	0.150	0.023	0.014	0.080
$g^\star$	−0.44	1.52	−14.9	−51.6	1.26
E_0 (eV)	1.52	3.13	0.42	0.24	1.42
Δ_0 (eV)	0.34	0.29	0.38	0.82	0.11
P (eVÅ)	10.49	8.97	9.20	9.64	8.85
E_0' (eV)	4.49	4.54	4.39	3.16	4.72
Δ_0' (eV)	0.17	0.15	0.24	0.33	0.07
P' (eVÅ)	4.78i	4.78i	0.87i	6.32i	2.87i
Q (eVÅ)	8.165	8.165	8.331	8.130	7.216

The constant of proportionality is a material-dependent spin–orbit coupling parameter. The x-, y-, and z-directions are chosen to be along (100), (010), and (001), respectively.

Conduction band effective mass revisited. Earlier we introduced the conduction band effective mass based on a very simple theory neglecting the spin–orbit interaction. A more elaborate theory including spin–orbit effects leads to an admixture of more bands than the heavy and light holes. As a consequence, more band edge parameters enter the expression (cf. Fig. 3.16 and Table 3.6). To a good approximation we have

$$\begin{aligned}\frac{m_\mathrm{e}}{m^\star} &= 1+\frac{1}{3}\frac{2m_\mathrm{e}P^2}{\hbar^2}\left(\frac{2}{E_0}+\frac{1}{E_0+\Delta_0}\right)\\ &\quad -\frac{1}{3}\frac{2m_\mathrm{e}P'^2}{\hbar^2}\left(\frac{2}{E_0'-E_0+\Delta_0'}+\frac{1}{E_0'-E_0}\right).\end{aligned}\tag{3.30}$$

Effective conduction band g-factor. It turns out that the spin–orbit interaction also affects the energy splitting of conduction band states in an external magnetic field B. This splitting is known as the Zeeman effect. The Zeeman energy splitting $\Delta E = g^\star\mu_\mathrm{B}B$ contains an effective g-factor which is material specific, whereas $g = 2$ for the free electron in vacuum. The spin–orbit interaction leads to a renormalization of the free-electron g. Some example values are tabulated in Table 3.6. The effective g-factor can be calculated from the band edge parameters given in the same table. It is, in good approximation, given by

$$\begin{aligned}g^\star &= 2-\frac{2}{3}\frac{2m_\mathrm{e}P^2}{\hbar^2}\left(\frac{1}{E_0}-\frac{1}{E_0+\Delta_0}\right)\\ &\quad +\frac{2}{3}\frac{2m_\mathrm{e}P'^2}{\hbar^2}\left(\frac{1}{E_0'-E_0}-\frac{1}{E_0'-E_0+\Delta_0'}\right).\end{aligned}\tag{3.31}$$

In the limit of vanishing spin–orbit interaction we have $\Delta_0 \to 0$ and $\Delta_0' \to 0$, and therefore $g^\star \to g = 2$. This means that the deviation of

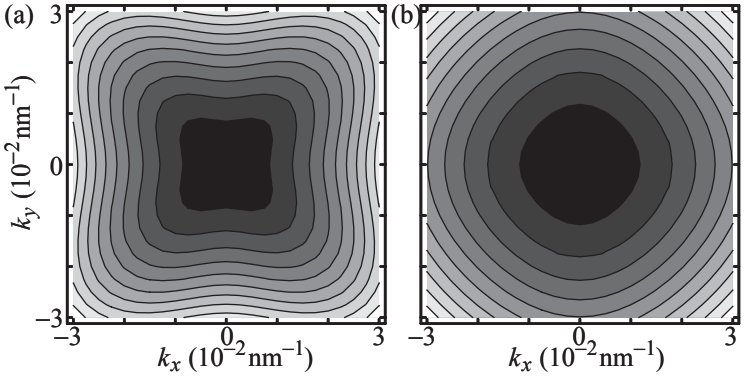


Fig. 3.17 Grayscale plots of (a) heavy and (b) light hole dispersion relations in the k_x-k_y plane calculated with the GaAs parameters. The grayscales for (a) and (b) are chosen to be different in order to emphasize the effect of warping.

the g-factor from the free-electron value is a result of spin–orbit interaction. The effective $g^\star$-factor becomes relevant in the next chapter, where the motion of crystal electrons in magnetic fields is considered [see in particular the effective-mass Schrödinger eq. (4.5)].

Dispersion relation for the valence band. The description of the valence band dispersion using an effective mass is problematic because there is a four-fold degeneracy at Γ. From $\mathbf{k}\cdot\mathbf{p}$-theory we can derive the following expression for the two dispersions for heavy and light holes (hh/lh) close to Γ

$$E_{\mathrm{hh/lh}} = -Ak^2 \mp \sqrt{B^2k^4 + |C|^2\left(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2\right)}, \qquad (3.32)$$

where the negative sign refers to the heavy hole dispersion. The material-specific parameters A, B, and C are tabulated for many semiconductors. The parameter C is responsible for the nonspherical warping of the valence band (see Fig. 3.17). This warping leads to different effective masses in different crystallographic directions. For example, in GaAs we have $A = -6.9$, $B = -4.4$, and $|C|^2 = 43$ (in units of $\hbar^2/2m_\mathrm{e}$). Figure 3.17 shows a grayscale plot of the heavy hole (a) and light hole (b) dispersion relations in the k_x-k_y plane. The parameters A, B, and C can be expressed with the band edge parameters from $\mathbf{k}\cdot\mathbf{p}$-theory:

$$\begin{aligned}
\frac{2m_\mathrm{e}}{\hbar^2}A &= 1 - \frac{2}{3}\left(\frac{P^2}{m_\mathrm{e}E_0} + \frac{2Q^2}{m_\mathrm{e}E_0'}\right)\\
\frac{2m_\mathrm{e}}{\hbar^2}B &= \frac{2}{3}\left(-\frac{P^2}{m_\mathrm{e}E_0} + \frac{Q^2}{m_\mathrm{e}E_0'}\right)\\
\left(\frac{2m_\mathrm{e}}{\hbar^2}C\right)^2 &= \frac{16P^2Q^2}{3m_\mathrm{e}E_0 m_\mathrm{e}E_0'}
\end{aligned}$$

It is often desirable and convenient to use an isotropic approximation of the form

$$E_{\mathrm{hh/lh}} = \frac{\hbar^2k^2}{2m_{\mathrm{hh/lh}}}$$

for the heavy and light hole dispersion relations. It can be obtained from eq. (3.32) by averaging over all directions in k-space. The result for the heavy and light hole masses is

$$\begin{aligned}\frac{1}{m_{\mathrm{hh}}} &= \frac{1}{\hbar^2}\left[-2A+2B\left(1+\frac{2|C|^2}{15B^2}\right)\right]\\ \frac{1}{m_{\mathrm{lh}}} &= \frac{1}{\hbar^2}\left[-2A-2B\left(1+\frac{2|C|^2}{15B^2}\right)\right].\end{aligned}$$

Constant energy surfaces for valence bands. There are two surfaces of constant energy near the valence band maximum at Γ, because of the existence of heavy and light hole bands. Surfaces of constant energy are strongly warped and deviate significantly from spheres (see Fig. 3.17), in contrast to the GaAs conduction band.

3.7 Thermal occupation of states

At zero temperature all valence band states are filled in a semiconductor without impurities and defects, while all conduction band states are empty. At finite temperatures, electrons can be thermally excited from the valence to the conduction band leading to the conduction band electron density

$$n_{\mathrm{c}}(T)=\int_{E_{\mathrm{c}}}^{\infty} dE\mathcal{D}_{\mathrm{c}}(E)\cdot\frac{1}{e^{(E-\mu)/k_{\mathrm{B}}T}+1}.$$

The density of the missing electrons in the valence band (i.e., the density of holes) is given by

$$\begin{aligned}p_{\mathrm{v}}(T) &= \int_{-\infty}^{E_{\mathrm{v}}} dE\mathcal{D}_{\mathrm{v}}(E)\left(1-\frac{1}{e^{(E-\mu)/k_{\mathrm{B}}T}+1}\right)\\ &= \int_{-\infty}^{E_{\mathrm{v}}} dE\mathcal{D}_{\mathrm{v}}(E)\frac{1}{e^{(\mu-E)/k_{\mathrm{B}}T}+1}.\end{aligned}$$

Fig. 3.18 shows the meaning of the individual factors below the integral for the density of the occupied states. A semiconductor in which the density of electrons and holes is governed by these relations is called an *intrinsic semiconductor*. This implies very low impurity and defect concentrations and no intentional doping.

Here we consider the case $E_{\mathrm{c}}-\mu \gg k_{\mathrm{B}}T$ and $\mu - E_v \gg k_{\mathrm{B}}T$ allowing the approximations

$$\begin{aligned}\frac{1}{e^{(E-\mu)/k_{\mathrm{B}}T}+1} &\approx e^{-(E-\mu)/k_{\mathrm{B}}T},\ \text{if } E>E_{\mathrm{c}}\\ \frac{1}{e^{(\mu-E)/k_{\mathrm{B}}T}+1} &\approx e^{-(\mu-E)/k_{\mathrm{B}}T},\ \text{if } E<E_{\mathrm{v}}.\end{aligned}$$

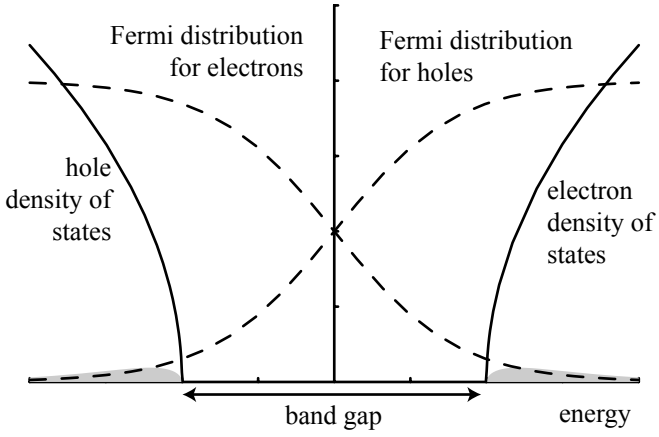


Fig. 3.18 Thermal occupation of electron and hole states in the conduction and valence band of a semiconductor. The carrier densities are thermally activated.

With these simplifications we obtain

$$\begin{aligned} n_{\mathrm{c}}(T) &= \int_{E_{\mathrm{c}}}^{\infty} dE \mathcal{D}_{\mathrm{c}}(E) e^{-(E-\mu)/k_{\mathrm{B}}T} \\ &= \underbrace{\int_{E_{\mathrm{c}}}^{\infty} dE \mathcal{D}_{\mathrm{c}}(E) e^{-(E-E_{\mathrm{c}})/k_{\mathrm{B}}T}}_{:=N_{\mathrm{c}}(T)} e^{-(E_{\mathrm{c}}-\mu)/k_{\mathrm{B}}T} \\ p_{\mathrm{v}}(T) &= \int_{-\infty}^{E_{\mathrm{v}}} dE \mathcal{D}_{\mathrm{v}}(E) e^{-(\mu-E)/k_{\mathrm{B}}T} \\ &= \underbrace{\int_{-\infty}^{E_{\mathrm{v}}} dE \mathcal{D}_{\mathrm{v}}(E) e^{-(E_{\mathrm{v}}-E)/k_{\mathrm{B}}T}}_{P_{\mathrm{v}}(T)} e^{-(\mu-E_{\mathrm{v}})/k_{\mathrm{B}}T} \end{aligned} \tag{3.33}$$

In an intrinsic semiconductor the densities are related via $n_{\mathrm{c}}(T) = p_{\mathrm{v}}(T)$, and we therefore obtain

$$\begin{aligned} n_{\mathrm{c}}(T) p_{\mathrm{v}}(T) = n_{\mathrm{c}}^2(T) &= N_{\mathrm{c}}(T) P_{\mathrm{v}}(T) e^{-(E_{\mathrm{c}}-E_{\mathrm{v}})/k_{\mathrm{B}}T} \\ &= N_{\mathrm{c}}(T) P_{\mathrm{v}}(T) e^{-E_{\mathrm{g}}/k_{\mathrm{B}}T} \end{aligned}$$

and as a consequence

$$n_{\mathrm{c}}(T) = \sqrt{N_{\mathrm{c}}(T) P_{\mathrm{v}}(T)} e^{-E_{\mathrm{g}}/2k_{\mathrm{B}}T}.$$

Comparing with eq. (3.33) allows the determination of the chemical potential μ as a function of temperature:

$$\mu(T) = E_{\mathrm{v}} + \frac{E_{\mathrm{g}}}{2} + \frac{k_{\mathrm{B}}T}{2} \ln \frac{P_{\mathrm{v}}(T)}{N_{\mathrm{c}}(T)}.$$

This equation implies that the chemical potential (Fermi level) is exactly in the middle of the band gap for $T \to 0$. Typically, the logarithmic ratio $\ln P_v(T)/N_c(T)$ will be of the order one and the Fermi level will not move away from the gap center by more than $k_{\mathrm{B}}T$. Therefore, our calculation is valid for $k_{\mathrm{B}}T \ll E_{\mathrm{g}}$, i.e. for most typical semiconductors even at room temperature ($k_{\mathrm{B}}T \approx 25$ meV).

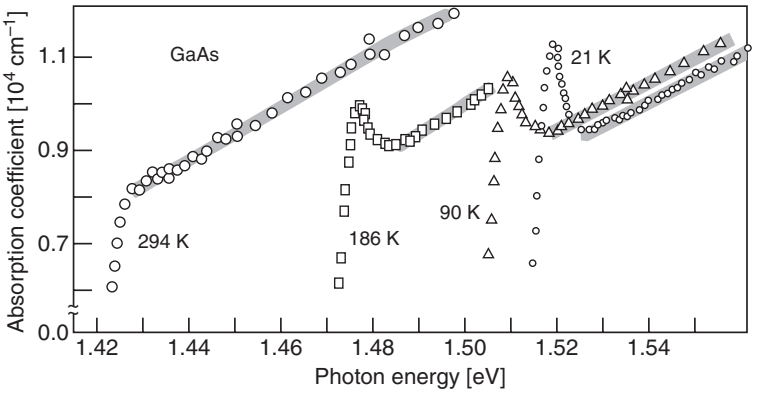


Fig. 3.19 Absorption coefficient of GaAs measured at different temperatures as a function of photon energy. At temperatures of 186 K and below, the exciton resonance can be seen. (Reprinted with permission from Sturge, 1962. Copyright 1962 by the American Physical Society. See also Yu and Cardona, 2001.)

3.8 Measurements of the band structure

Interband absorption and emission. Semiconductors can absorb and emit photons involving electronic transitions between valence and conduction band states. These optical transitions obey energy and momentum conservation. The energy of the photon $h\nu$ has to be equal to the energetic difference of the participating valence and conduction band state:

$$h\nu = E_{\mathrm{c}}(\mathbf{k}_{\mathrm{c}}) - E_{\mathrm{v}}(\mathbf{k}_{\mathrm{v}}).$$

The photon momentum $\hbar\mathbf{k}$ has to be taken up (or released) by the electronic system, meaning that

$$\hbar\mathbf{k} = \hbar\mathbf{k}_{\mathrm{c}} - \hbar\mathbf{k}_{\mathrm{v}}.$$

For absorption to take place, the energies of photons have to exceed the band gap, i.e., depending on the material it must be in the wavelength range of a few microns (infrared) up to a few hundred nanometers (visible region; $E = 1.24\ \mathrm{eV}\,\mu\mathrm{m}/\lambda$). This range of wavelengths is suitable for optical applications, such as light emitting diodes or semiconductor lasers. The momentum of a photon at these energies is of the order of $5 \times 10^4\ \mathrm{cm}^{-1}$, which is very small compared to the wave vectors of the electrons $\pi/a \approx 5 \times 10^7\ \mathrm{cm}^{-1}$. Therefore we talk about vertical transitions between $E(\mathbf{k})$ bands.

Absorption measurements near the fundamental gap are therefore a suitable method for determining the fundamental gap of a semiconductor experimentally. Figure 3.19 shows the absorption coefficient of GaAs measured at various temperatures. At room temperature, optical absorption sets in slightly above 1.42 eV, i.e., as soon as the photon energy exceeds the band gap. Below this energy, the semiconductor is transparent. At lower temperatures the absorption edge is shifted to higher energies (blue shift). Qualitatively this is a result of the thermal contraction of the crystal lattice leading to a slightly reduced lattice constant and thereby to an increased band gap. In addition, a strongly enhanced absorption is seen at the absorption edge at low temperatures. This

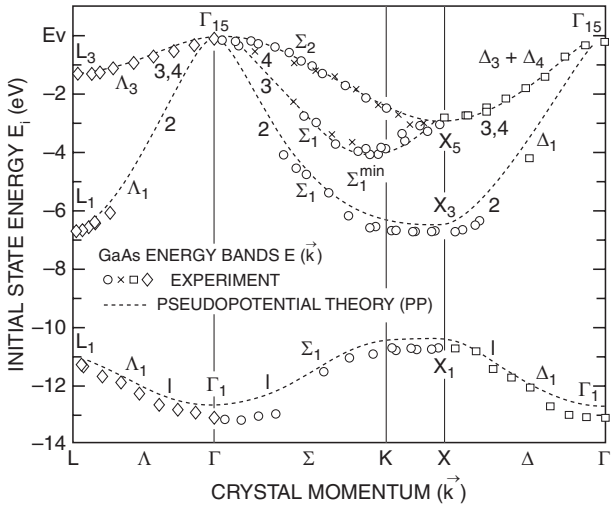


Fig. 3.20 Measured and calculated valence band structure of GaAs. The calculations were performed using the pseudopotential method. Measurements were made using angle-resolved photoemission. (Reprinted with permission from Chiang *et al.*, 1980. Copyright 1980 by the American Physical Society.)

is the so-called exciton resonance. The optical excitation creates an electron–hole pair consisting of a positively charged hole and a negatively charged electron (see also section 4.2). Their mutual Coulomb interaction leads to a bound state with typical binding energies of about 5 meV. Such an interacting electron–hole pair is called an *exciton*. Note that, typically, the electron and the hole have different effective masses.

Angle resolved photoemission spectroscopy (ARPES). The band structure of the valence band can be measured with angle-resolved photoemission. This technique uses photons with an energy large enough to extract electrons from the crystal (photoeffect). Synchrotron radiation experiments use photons in the range between 25 and 100 eV, such that the final state of the electron after photon absorption is far beyond the work function ($|e|\Phi = 5.15$ eV for GaAs), i.e., at an energy where the influence of the lattice potential is negligibly small and the electrons can be described with free electron states.

Figure 3.20 shows the valence band structure of GaAs measured in this way in comparison to a pseudopotential calculation. Experiment and calculation agree very well. The experiment was performed on a [110] oriented GaAs sample with the photons incident onto the surface at an angle of 45°. In this case, the wave vectors of the extracted electrons are given by

$$\hbar k_{\parallel} = \sqrt{2m_{\mathrm{e}}(E_i + h\nu - e\Phi)}\,\sin\theta$$

and

$$\hbar k_{\perp} = \sqrt{2m_{\mathrm{e}}[(E_i + h\nu - e\Phi)\cos^2\theta - V_0]},$$

where θ is the emission angle and $V_0 = -14.5$ eV is the bottom of the 'muffin tin' potential measured from the vacuum level. For a certain direction of $\mathbf{k}$, the initial energy E_i and the magnitude k of the wave vector can be determined from these two equations, because the photon

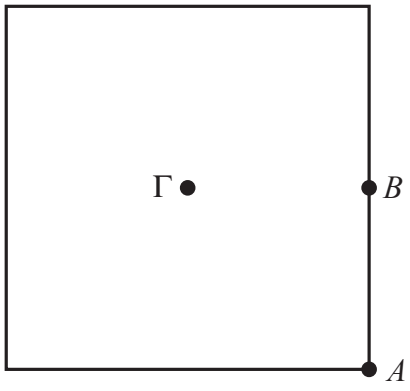
energy $h\nu$ and the work function $e\Phi$ are known, and the angle θ can be chosen.

Further reading

- Band structure, general introduction: Kittel 2005; Ashcroft and Mermin 1987; Singleton 2001.
- Band structure of semiconductors: Yu and Cardona 2001; Cohen and Chelikowski 1989; Balkanski and Wallis 2000.
- Density of states and thermal occupation: Ashcroft and Mermin 1987; Seeger 2004.
- Graphene and carbon nanotubes: Ando 2005.

Exercises

(3.1) Sketch the free electron band structure of a two-dimensional square lattice along the path Γ-*A*-*B*-Γ. Write down the wave functions with the lowest energies at these three points.



(3.2) In this problem we try to approximate the π-band dispersion relation of graphene (see page 23). To this end we describe the hexagonal two-dimensional crystal lattice (see Fig. 3.5) as a Bravais lattice with the two basis vectors

$$\begin{aligned} \mathbf{a}_1 &= \sqrt{3}a_0(1/2, \sqrt{3}/2) \\ \mathbf{a}_2 &= \sqrt{3}a_0(-1/2, \sqrt{3}/2), \end{aligned}$$

where a_0 is the nearest neighbor separation. The primitive cell spanned by these two vectors contains the basis of the lattice consisting of two atoms A and B (see Fig 3.5).

(a) Determine the basis vectors of the reciprocal lattice and construct the first Brillouin zone (BZ).

(b) Consider the model of free electrons with the dispersion in eq. (3.10). Draw the dispersion relations along lines leading from the center of the first BZ to one of its corners (K-point), and to the center of its boundary lines (M-point). Restrict this drawing to the two sets of shortest reciprocal lattice vectors. Compare your drawing to Fig. 3.7.

(c) Consider the lowest doubly degenerate states at the M-point on the boundary of the first BZ. Calculate the splitting of these degenerate states under the influence of a small potential modulation.

(d) Find reasons why the periodic potential modulation leaves a two-fold degeneracy at the lowest K-point energy.

(3.3) Draw the first Brillouin zone of an fcc lattice as seen in the k_z-direction.

(3.4) Implement a pseudopotential bandstructure calculation in Mathematica (or another program of

you choice) for determining the band structure of diamond and zincblende semiconductors approximately.

(3.5) Given the band gaps of InAs (0.36 eV) and InP (1.27 eV) estimate the relative effective masses $m^\star/m$ of conduction band electrons using eq. (3.21). Compare with the values tabulated in Table 3.5.

(3.6) Calculate the analytical expression for the density of states of two-, and one-dimensional systems assuming a two-, or one-dimensional parabolic dispersion relation.

(3.7) Calculate an analytical expression for the density of states near the conduction band minimum of silicon using the dispersion relation in eq. (3.24).

(3.8) Consider a gas containing N electrons. The change dE in total energy of the system in response to a small change in entropy (dS) in volume (dV) or in electron number dN is given by

$$dE(S,V,N) = TdS + pdV + \mu_{\rm ch}dN + UdQ,$$

where $\mu_{\rm ch}$ is the chemical potential, U is the electrostatic potential and $dQ = -edN$ is the change of charge in the system upon a change in electron number.

(a) Discuss, why, in this system of charged particles, the electrochemical potential $\mu = \mu_{\rm ch} - eU$ is a state variable. (In general, U could even vary spatially.)

(b) Let this system be isolated such that its total energy E, its total volume V and its electron number N cannot change. In a *Gedankenexperiment* split the system into two subsystems 1 and 2 which obey

$$\begin{aligned} E_1 + E_2 &= E = \text{const.} \\ V_1 + V_2 &= V = \text{const.} \\ N_1 + N_2 &= N = \text{const.} \end{aligned}$$

Show using the second law of thermodynamics that the two systems will have the same temperature $T_1 = T_2$, the same pressure $p_1 = p_2$, and the same electrochemical potential $\mu_1 = \mu_2$.

Envelope functions and effective mass approximation

4

In this chapter we are interested in the quantum mechanical motion of electrons in the crystal if the periodic lattice potential is perturbed. This can occur as a result of the presence of lattice defects, impurities, or doping atoms. It can also arise due to the incorporation of interfaces between different layers of materials. Other reasons could be the presence of external electric or magnetic fields, or internal fields arising from time-dependent lattice distortions or vibrations such as those caused by phonons or surface acoustic waves. In this chapter, we will restrict ourselves to static perturbations small enough to be treated in lowest order perturbation theory, and of a spatial range much larger than the lattice constant of the underlying material. We will see that this restriction leads to considerable simplifications leading us to an effective mass Schrödinger equation for electrons in conduction bands with parabolic dispersion.

4.1 Quantum mechanical motion in a parabolic band

Weak and long-range perturbations of perfect crystal symmetry can be caused, for example, by an external electric field, or by the presence of a charged doping atom. Figure 4.1 shows schematically the perturbed

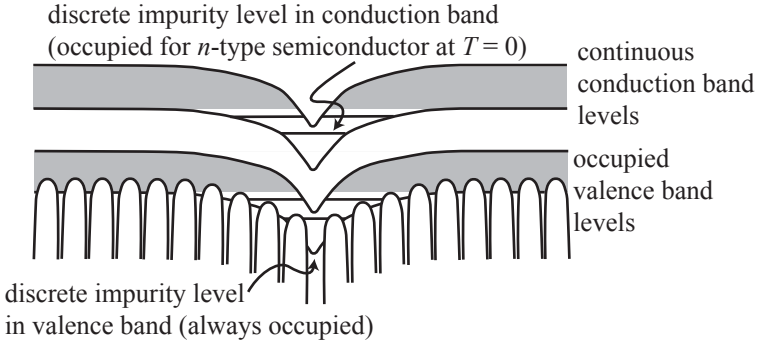


Fig. 4.1 Continuum and discrete energy levels in the vicinity of a doping atom in a semiconductor. E_1 is the energy of a discrete level below the conduction band edge; E_2 is the energy of a state in the continuum. (Reprinted with permission from Slater, 1949. Copyright 1949 by the American Physical Society.)

lattice potential in the presence of a positively charged doping atom.

There are a number of different ways of solving this quantum mechanical problem for the electronic motion. The methods differ essentially in the set of basis functions used as a starting point for a perturbation treatment. People have used Bloch-states (Enderlein and Schenk, 1992), band edge states from $\mathbf{k}\cdot\mathbf{p}$-theory (Luttinger and Kohn, 1955), and the so-called Wannier states (Wannier, 1937; Zinman, 1972; Kittel, 1970). In order to give some insight into the derivation of the equation of motion, we will work in the Bloch-state basis and restrict the discussion to a perturbation of a parabolic conduction band with minimum at Γ as it is found, for example, in GaAs.

The problem on the basis of Bloch-states. Assume that we have solved Schrödinger's equation for the unperturbed crystal. The corresponding dispersion relations $E_n(\mathbf{k})$ and the Bloch-functions $\psi_{n\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{kr}}u_{n\mathbf{k}}(\mathbf{r})$ are known. Now we seek the solution of the perturbed Schrödinger equation

$$[H_0 + U(\mathbf{r})]\,\Psi(\mathbf{r}) = E\Psi(\mathbf{r}), \tag{4.1}$$

where H_0 is the hamiltonian of the unperturbed lattice and $U(\mathbf{r})$ is the perturbing potential. We expand the wave function $\Psi(\mathbf{r})$ on the basis of Bloch-states:

$$\Psi(\mathbf{r}) = \sum_{n,\mathbf{k}} F_n(\mathbf{k})\psi_{n\mathbf{k}}(\mathbf{r}).$$

Inserting this expansion into Schrödinger's equation gives

$$\sum_{n\mathbf{k}} \psi_{n\mathbf{k}}(\mathbf{r})\,[E_n(\mathbf{k}) - E + U(\mathbf{r})]\,F_n(\mathbf{k}) = 0.$$

Multiplying by $\psi^\star_{n'\mathbf{k}'}(\mathbf{r})$ and integrating over $\mathbf{r}$ leads to

$$\sum_{n,\mathbf{k}} [(E_n(\mathbf{k}) - E)\,\delta_{n\mathbf{k},n'\mathbf{k}'} + U_{n'\mathbf{k}',n\mathbf{k}}]\,F_n(\mathbf{k}) = 0, \tag{4.2}$$

where we have used the orthogonality of Bloch-states and introduced the matrix elements of the perturbing potential

$$U_{n'\mathbf{k}',n\mathbf{k}} = \int d^3r\ \ \psi^\star_{n'\mathbf{k}'}(\mathbf{r})U(\mathbf{r})\psi_{n\mathbf{k}}(\mathbf{r}).$$

The matrix elements of the perturbation. We will now further simplify the matrix elements of the perturbation. To this end we introduce the Fourier transform of $U(\mathbf{r})$ (see Appendix A.2) and obtain

$$U_{n'\mathbf{k}',n\mathbf{k}} = \int d^3q\ \ U(\mathbf{q}) \int d^3r\ \ e^{i(\mathbf{k}-\mathbf{k}'+\mathbf{q})\mathbf{r}} u^\star_{n'\mathbf{k}'}(\mathbf{r})u_{n\mathbf{k}}(\mathbf{r}).$$

In this expression we can expand the lattice periodic function $u^\star_{n'\mathbf{k}'}(\mathbf{r})u_{n\mathbf{k}}(\mathbf{r})$ into a Fourier series and obtain for the matrix element

$$U_{n'\mathbf{k}',n\mathbf{k}} = \int d^3q\ \ U(\mathbf{q}) \sum_{\mathbf{K}} C^{n'\mathbf{k}'}_{n\mathbf{k}}(\mathbf{K}) \int d^3r\ \ e^{i(\mathbf{k}-\mathbf{k}'+\mathbf{q}+\mathbf{K})\mathbf{r}}$$

with the so-called Bloch integral

$$C_{n\mathbf{k}}^{n'\mathbf{k}'}(\mathbf{K}) = \frac{1}{V_0}\int_{EZ} d^3r \;\; e^{-i\mathbf{K}\mathbf{r}} u^{\star}_{n'\mathbf{k}'}(\mathbf{r}) u_{n\mathbf{k}}(\mathbf{r}).$$

The spatial integral in the expression for the matrix element $U_{n'\mathbf{k}',n\mathbf{k}}$ contributes only if the exponent vanishes, i.e., if $\mathbf{q} = \mathbf{k}' - \mathbf{k} - \mathbf{K}$. As a matter of fact, the integral is a representation of Dirac's delta function. Therefore the matrix element simplifies to

$$U_{n'\mathbf{k}',n\mathbf{k}} = (2\pi)^3 \sum_{\mathbf{K}} U(\mathbf{k}' - \mathbf{k} - \mathbf{K}) C_{n\mathbf{k}}^{n'\mathbf{k}'}(\mathbf{K}). \tag{4.3}$$

So far we have used the periodicity of the crystal lattice without using any approximation.

Simplifying approximations. For further simplifications to the problem we make the following assumptions about the perturbation:

(1) We assume that the perturbing potential changes slowly on the scale of the lattice constant, i.e., $U(\mathbf{q})$ is significant only for $q \ll \pi/a$.

(2) We assume that the perturbation is small compared to typical energy separations of bands in the crystal.

(3) We assume that the coefficients $F_n(\mathbf{k})$ have significant values only for small values of $\mathbf{k}$.

According to the third assumption, we consider only states near the nondegenerate Γ-minimum. As a consequence of this and the first assumption, in the sum over $\mathbf{K}$ only $\mathbf{K} = 0$ is retained and the matrix element simplifies to

$$U_{n'\mathbf{k}',n\mathbf{k}} \approx (2\pi)^3 U(\mathbf{k}' - \mathbf{k}) C_{n\mathbf{k}}^{n'\mathbf{k}'}(0).$$

Now we would like to simplify the Bloch integral $C_{n\mathbf{k}}^{n'\mathbf{k}'}(0)$. Based on the third assumption, we employ the expansion of the Bloch-functions near the conduction band minimum, eq. (3.19). We obtain

$$C_{n\mathbf{k}}^{n'\mathbf{k}'}(0) = \frac{1}{V_0}\int_{EZ} d^3r \;\; u^{\star}_{n'\mathbf{k}'}(\mathbf{r}) u_{n\mathbf{k}}(\mathbf{r}) \approx \frac{1}{(2\pi)^3}\delta_{nn'} + \mathcal{O}(k^2),$$

and therefore

$$U_{n'\mathbf{k}',n\mathbf{k}} \approx U(\mathbf{k}' - \mathbf{k})\delta_{nn'}.$$

This means that, given our assumptions, the perturbation does not mix states of neighboring bands, but only states of different $\mathbf{k}$ near the Γ-minimum. With the above result for the matrix element, the equation of motion (4.2) simplifies to

$$\sum_{\mathbf{k}} \left[\left(E_n(\mathbf{k}) - E\right) \delta_{\mathbf{k},\mathbf{k}'} + U(\mathbf{k}' - \mathbf{k}) \right] F_n(\mathbf{k}) = 0.$$

Simplification of the wave function. The wave function in real space now reads

$$\Psi(\mathbf{r}) = \sum_{\mathbf{k}} F_n(\mathbf{k}) e^{i\mathbf{kr}} u_{n\mathbf{k}}(\mathbf{r}).$$

Only small wave vectors $\mathbf{k}$ are important here, due to the long-range nature of $U(\mathbf{r})$. We therefore approximate $u_{n\mathbf{k}}(\mathbf{r}) \approx u_{n0}(\mathbf{r})$ and obtain for the wave function

$$\Psi(\mathbf{r}) = u_{n0}(\mathbf{r}) \sum_{\mathbf{k}} F_n(\mathbf{k}) e^{i\mathbf{kr}} = u_{n0}(\mathbf{r}) F_n(\mathbf{r}).$$

In the last step we have interpreted the sum over $\mathbf{k}$ as the Fourier series of a real space function $F_n(\mathbf{r})$. This function is of long range compared to the lattice period and is called the *envelope function* of the wave function.

Approximating the dispersion. We now approximate the dispersion relation $E_n(\mathbf{k})$ accordingly by using an approximation for small $\mathbf{k}$. Near the Γ-minimum we have [cf. eq. (3.22)]

$$E_{\mathrm{c}}(\mathbf{k}) = E_{\mathrm{c}} + \frac{\hbar^2 k^2}{2m^\star},$$

where $m^\star$ is the effective mass of electrons in the conduction band. With these simplifications the equation of motion for electrons reads

$$\frac{\hbar^2}{2m^\star} k^2 F_{\mathrm{c}}(\mathbf{k}) + \sum_{\mathbf{k}'} U(\mathbf{k}-\mathbf{k}') F_{\mathrm{c}}(\mathbf{k}') = (E - E_{\mathrm{c}}) F_{\mathrm{c}}(\mathbf{k}).$$

Equation of motion in real space. This equation determines the Fourier components of the envelope function $F_{\mathrm{c}}(\mathbf{r})$. Transformation from Fourier space into real space is straightforward. The first term on the left-hand side corresponds to the second derivative of the envelope function in real space. The second term is a convolution integral which transforms into the product of the two corresponding functions in real space. We therefore obtain the following differential equation determining the envelope function $F_{\mathrm{c}}(\mathbf{r})$:

$$\Big[-\frac{\hbar^2}{2m^\star}\Delta + \underbrace{E_{\mathrm{c}} + U(\mathbf{r})}_{:=E_{\mathrm{c}}(\mathbf{r})} \Big] F_{\mathrm{c}}(\mathbf{r}) = E F_{\mathrm{c}}(\mathbf{r}). \tag{4.4}$$

This is exactly Schrödinger's equation (4.1) where the periodic lattice potential hidden in H_0 has disappeared, but the free electron mass in H_0 has been replaced by the effective mass of the conduction band electrons. Introducing the local band edge energy $E_{\mathrm{c}}(\mathbf{r})$, this function acts as the effective potential in which the conduction band electrons move.

The envelope function $F_{\mathrm{c}}(\mathbf{r})$ brings about very convenient simplifications. For example, matrix elements of a quantum mechanical quantity, which have to be calculated using the complete electronic wave function,

can usually be expressed as integrals over the envelope function alone. As an example, we consider the electron density. Assume that the envelope functions $F_i(\mathbf{r})$ are solutions of eq. (4.4) with energies E_i. The electron density of the system is then given by

$$n(\mathbf{r}) = \sum_i |\psi_i(\mathbf{r})|^2 f(E_i) = |u_{c0}(\mathbf{r})|^2 \sum_i |F_i(\mathbf{r})|^2 f(E_i),$$

where $f(E)$ is the Fermi distribution function. The envelope function and the lattice periodic function $u_{c0}(\mathbf{r})$ vary on different length scales. Within a primitive cell at position $\mathbf{R}$ of the lattice $F_i(\mathbf{r}) \approx F_i(\mathbf{R})$ is essentially constant. If we are interested only in the mean density in the primitive cell at $\mathbf{R}$, it is given by

$$n(\mathbf{R}) = \underbrace{\frac{1}{V_{EZ}} \int_{EZ} dV\, |u_{c0}(\mathbf{r})|^2}_{=1} \sum_i |F_i(\mathbf{R})|^2 f(E_i) = \sum_i |F_i(\mathbf{R})|^2 f(E_i).$$

On a length scale that is large compared to the lattice constant, the electron density is given by the envelope function alone and we can neglect the lattice periodic function $u_{n0}(\mathbf{r})$.

Hydrogen-like impurities. A simple application of the concept of the envelope function is the determination of the energy levels of a hydrogen-like impurity in a semiconductor. It has indeed been shown that modern fabrication techniques have the potential to allow a precise incorporation of single doping atoms at predefined locations. Figure 4.2 shows scanning tunneling microscope images of a hydrogen passivated Si(001) surface. Using the tip of the scanning tunneling microscope, hydrogen atoms can be locally desorbed. Such a spot of about 1 nm size is shown in Fig. 4.2(a). If the surface is then exposed to PH_3, the molecules are preferentially adsorbed at those positions, where the hydrogen passivation has been removed. A thermal annealing step lets the P atom diffuse into the top layer of the Si substrate where it forms a substitutional doping site as shown in Fig. 4.2(b).

As an example for the use of the effective mass equation, we consider a silicon atom sitting on the Ga site in a GaAs lattice. The silicon atom can satisfy all bonds with neighboring arsenic atoms using only three of its four valence electrons. As a consequence, one excess electron and an excess positive elementary charge in the silicon nucleus remain. Such a silicon atom is called a donor, because it can give away the excess electron. However, the positively charged donor ion will bind the excess electron, and the Coulomb interaction between them will appear in the equation for the envelope function:

$$\left[-\frac{\hbar^2}{2m^\star} \Delta - \frac{e^2}{4\pi\varepsilon\varepsilon_0 r} \right] F_c(\mathbf{r}) = (E - E_c) F_c(\mathbf{r}).$$

The important point is that the relative dielectric constant of the host crystal, in our case GaAs, enters in the Coulomb potential. It accounts

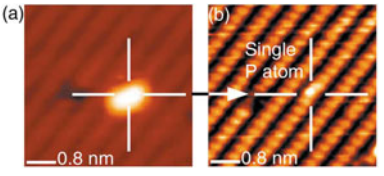


Fig. 4.2 STM images of atomically controlled single phosphor atom incorporation into Si(001). (a) Hydrogen terminated Si(001) surface with a hydrogen desorption point. (b) The same area after PH_3 dosing and annealing showing a single P atom incorporated at the location defined by the H-desorption point. (Reprinted with permission from Schofield, 2003. Copyright 2003 by the American Physical Society.)

for the polarization of the lattice by the charged donor, which effectively reduces the interaction strength. The solution of this quantum problem is that of the hydrogen problem, in which the Rydberg energy $E_{\rm Ry} = 13.6$ eV is replaced by an effective Rydberg energy $E^\star_{\rm Ry}$ and Bohr's radius $a_{\rm B} = 0.53$ Å by an effective radius $a^\star_{\rm B}$:

$$\begin{aligned} E^\star_{\rm Ry} &= \frac{e^4 m^\star}{2(4\pi\varepsilon\varepsilon_0)^2\hbar^2} = E_{Ry}\frac{m^\star}{m_{\rm e}}\frac{1}{\varepsilon^2} \\ a^\star_{\rm B} &= \frac{4\pi\varepsilon\varepsilon_0\hbar^2}{m^\star e^2} = a_{\rm B}\frac{m_{\rm e}}{m^\star}\varepsilon. \end{aligned}$$

For GaAs, with $\varepsilon = 12.53$ and $m^\star = 0.067 m_{\rm e}$, we find $E^\star_{\rm Ry} = 5.7$ meV and $a^\star_{\rm B} = 100$ Å. The energy levels of the hydrogen-like impurity are then

$$E_n = E_{\rm c} - \frac{E^\star_{\rm Ry}}{n^2}.$$

These states are discrete and lie below the conduction band edge of the unperturbed crystal as schematically shown in Fig. 4.3. As in the hydrogen atom, the excitation energy $E^\star_{\rm Ry}$ from the ground state to the lower edge of the conduction band (continuum) is called the binding energy. Measured binding energies of donors in GaAs are 5.789 meV for $\rm Se_{As}$, 5.839 meV for $\rm Si_{Ga}$, 5.870 meV for $\rm S_{As}$, 5.882 meV for $\rm Ge_{Ga}$, and 5.913 meV for $\rm C_{Ga}$. These values agree quite well with the theoretical prediction for $E^\star_{\rm Ry}$.

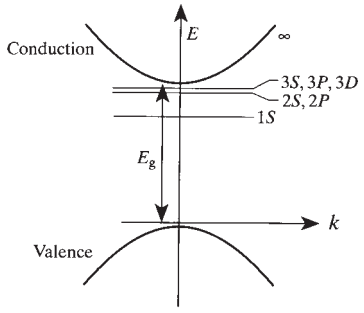


Fig. 4.3 Energy levels of a hydrogen-like impurity in GaAs (Yu and Cardona, 2001).

Figure 4.4 shows the total wave function of the ground state including the Bloch part emphasizing that the envelope function determines the shape of the probability density distribution on length scales large compared to the lattice constant.

Equation of motion at the Γ-minimum of the conduction band in the presence of a magnetic field. The equation of motion of an electron at the conduction band minimum under the influence of a

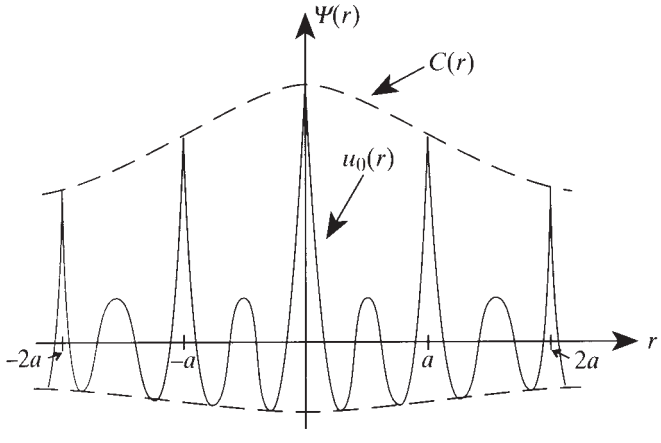


Fig. 4.4 Total wave function of the hydrogen-like impurity in GaAs including the Bloch contribution (Yu and Cardona, 2001).

magnetic field has been derived by Luttinger (1951), and by Luttinger and Kohn (1955) using similar methods. It was also found that, in this case, the equation for the envelope function is identical to the effective mass Schrödinger equation for a free particle in a magnetic field. Under the simultaneous influence of a vector potential $\mathbf{A}(\mathbf{r})$ and an electrostatic potential $U(\mathbf{r})$ the equation of motion for electrons at the Γ-minimum of the conduction band (see, e.g., Winkler 2003) reads

$$\left[\frac{1}{2m^\star}\left(\frac{\hbar}{i}\nabla + |e|\mathbf{A}(\mathbf{r})\right)^2 + U(\mathbf{r}) + \frac{1}{2}g^\star\mu_\mathrm{B}\boldsymbol{\sigma}\mathbf{B}\right]F_\mathrm{c}(\mathbf{r}) = (E - E_\mathrm{c})F_\mathrm{c}(\mathbf{r}). \tag{4.5}$$

Here, the elementary charge $|e| = 1.6 \times 10^{-19}\,\mathrm{C}$ is taken to be a positive number. In the following chapters of the book we will frequently call the envelope function $F_\mathrm{c}(\mathbf{r})$ simply the *wave function* of the electron, because its equation of motion is identical with that of an electron with mass $m^\star$ in vacuum. We will further use the convention that all energies are measured from the conduction band edge of the unperturbed crystal, such that $E_\mathrm{c} = 0$ in the above equation. The effective mass $m^\star$ and the effective $g^\star$-factor entering in the above equation can be calculated from the knowledge of the band edge parameters given in Table 3.6 using eqs. (3.30) and (3.31).

Equation (4.5) is of great importance for semiconductor nanostructures. Methods of structuring and patterning materials allow the fabrication of tailored potential landscapes $U(\mathbf{r})$. Magnetic fields can be created in the laboratory that influence the electronic motion as they do in the free electron case. Solving the equations of motion is greatly facilitated by the existence of many analytical solutions and approximative schemes from quantum mechanics textbooks.

The considerations leading to eq. (4.5) for conduction band electrons near Γ can be extended to semiconductors with conduction band minima at other points in the first Brillouin zone (e.g., silicon or germanium). In this case, the wave function is expanded at the corresponding conduction band minima rather than at Γ. More complicated equations of motion result due to the valley degeneracy and the anisotropic effective masses. The theory for valence band holes is also much more demanding, because there are degenerate states at Γ.

4.2 Semiclassical equations of motion, electrons and holes

Conduction band electrons. With the validity of the effective mass Schrödinger equation (4.5) for the crystal electron, the semiclassical limit of quantum mechanics (i.e., the motion of wave packets) must have its range of application in semiconductor physics. Wave packets can be constructed from the envelope functions $F_\mathrm{c}(\mathbf{r})$ and the dynamics of its center of mass can be investigated. The result is Newton's equation of

motion

$$m^\star \ddot{\mathbf{r}} = -|e|(\mathbf{E} - \dot{\mathbf{r}} \times \mathbf{B}), \tag{4.6}$$

where $\mathbf{E}$ is the electric field and $\mathbf{B}$ is the magnetic field at the location of the electron. As a consequence, there is a variety of possibilities in the physics of semiconductor nanostructures to investigate the borderlines between classical and quantum physics. Examples are investigations of the relation between classical and quantum chaos, or the transition from quantum to classical mechanics in the presence of decoherence.

Valence band holes. We will now briefly discuss the dynamics of holes, i.e., missing electrons near a maximum of the valence band, in the classical limit. The convex curvature of the valence band could be interpreted using a negative effective mass. Newton's equation of motion reads in this case

$$-m^\star \ddot{\mathbf{r}} = -|e|(\mathbf{E} - \dot{\mathbf{r}} \times \mathbf{B}).$$

However, a negative effective mass is physically not very intuitive. We can reinterpret this equation of motion by multiplying it by -1:

$$m^\star \ddot{\mathbf{r}} = +|e|(\mathbf{E} - \dot{\mathbf{r}} \times \mathbf{B})$$

This can be interpreted as the equation of motion for particles with positive mass $m^\star$, but with *positive charge* $+|e|$. The occurrence of a positive charge at the top of the valence band is also intuitive from another point of view. In the electrically neutral, uncharged semiconductor crystal the valence band is completely filled. Removing an electron from the top of the valence band, an initially localized positive charge remains. Such a missing electron is called a hole. According to the above equation of motion, the effective mass $m^\star$ and the charge $+e$ are properties of this hole which appears to move through the crystal like a classical particle.

Further reading

- Papers: Slater 1949; Luttinger 1951; Luttinger and Kohn 1955.
- Effective mass from $\mathbf{k} \cdot \mathbf{p}$-theory: Davies 1998; Kittel 1970; Yu and Cardona 2001.
- Effective mass from quasi-classical considerations with group velocity and Newton's equation of motion: Kittel 2005; Kittel 1970; Singleton 2001; Ashcroft and Mermin 1987.
- Effective mass from the hydrogen problem in semiconductors, doping: Davies 1998.
- Band structure of semiconductors: Winkler 2003.

Exercises

(4.1) Consider the differential equation for the envelope function, eq. (4.5), with a magnetic field $\mathbf{B} = (0, 0, B)$ and the Coulomb potential $U(\mathbf{r}) = e^2/4\pi\epsilon\epsilon_0 r$.

(a) Give reasons why the solution of the problem can be separated in that of the orbital motion and that of the spin dynamics.

(b) Discuss qualitatively the effects of the magnetic field on the spin dynamics.

(c) Discuss qualitatively how the magnetic field affects the orbital energy levels and wave functions.

(4.2) In silicon, the hamiltonian for the conduction band envelope function in the effective mass equation is given by

$$H = \frac{\hbar^2}{2m_\mathrm{L}}\frac{\partial^2}{\partial x^2} + \frac{\hbar^2}{2m_\mathrm{T}}\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + V_\mathrm{c}(r),$$

where $V_\mathrm{c} = e^2/4\pi\epsilon\epsilon_0 r$ is the Coulomb potential, and m_L and m_T are the longitudinal and transverse effective masses, respectively. Consider the case $m_\mathrm{L} = m_\mathrm{T} + \Delta m$, where $\Delta m/m_\mathrm{T} \ll 1$. Calculate the effect of the presence of Δm on the energies of the 1s-, 2s-, and 2p-states of a hydrogen-like impurity using perturbation theory.

This page intentionally left blank

Material aspects of heterostructures, doping, surfaces, and gating

5

5.1 Band engineering

The possibility of growing materials of very different composition atomic layer by atomic layer with molecular beam epitaxy provides a method of varying the band structure in the growth direction and tailoring it according to the requirements of electronic or optical devices.

Material aspects: Figure 5.1 shows the relation between lattice constant and band gap for a number of common binary semiconductor materials. Only materials of the same lattice constant and crystal structure can be grown on top of each other without creating strain. For example, a very common combination is the GaAs/AlAs material system.

In contrast, if, for example, InAs is grown on GaAs, a strained layer is formed due to the lattice mismatch. Experience shows that for this particular material combination InAs layers of up to 15 nm thickness can be grown before dislocations form that relax the strain. If an InAs layer below 15 nm thickness is sandwiched between thick GaAs layers, the lattice constant of this thin layer is strained such that it almost matches the GaAs lattice constant. Such a strained layer is called *pseudomorphic*.

Another possibility for combining layers of different materials is the combination of ternary alloys. In Fig. 5.1, the lines connecting two binary compounds having one constituent in common give the lattice constant–band gap relation of the corresponding ternary compound. As an example, we find that $In_{0.53}Ga_{0.47}As$ and $In_{0.52}Al_{0.48}As$ have the same lattice constant, but different band gaps.

In case of GaAs and AlAs the ternary compound $Ga_xAl_{1-x}As$ can be formed which has been extensively studied. Layers of GaAs can be combined with layers of $Ga_xAl_{1-x}As$ for arbitrary values of the Ga fraction x, because the lattice mismatch between GaAs and AlAs is only 0.15%. Figure 5.2 shows how the band gap of $Al_xGa_{1-x}As$ changes as a function of x. For $x < 0.45$ the band gap is direct; above it is indirect. Usually one grows material with $x < 0.4$ in order to make sure that the lowest conduction band minimum is at Γ. In this range of x the band

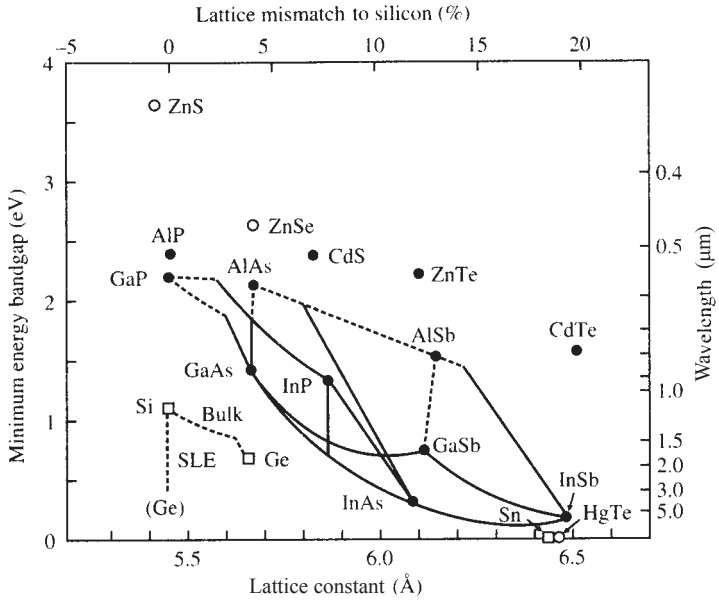


Fig. 5.1 Band gap of some common semiconductors vs lattice constant. The curves connect binary compounds that have one constituent in common, such as GaAs and AlAs mixing to $Al_xGa_{1-x}As$. Solid lines represent direct, dashed lines indirect band gaps (Singleton, 2001).

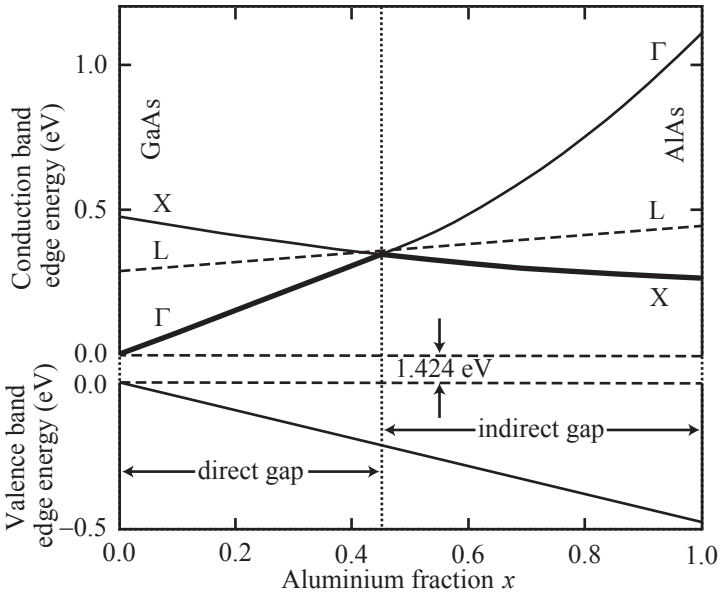


Fig. 5.2 Energies of the conduction band minimum at Γ, X, and L and the Γ-maximum of the valence band of $AI_xGa_{1-x}As$ at room temperature as a function of the aluminium fraction x. (Redrawn from Yu *et al.*, 1992 and Adachi, 1985.)

gap is given by

$$E_g = (1.516 + 1.247x)\ \text{eV for } Al_xGa_{1-x}As \text{ with } x < 0.45. \qquad (5.1)$$

This formula is valid for temperatures below one Kelvin.

Band edges at interfaces between different materials. With these possibilities for combining different materials, the question arises of how the band structure is changed at the interface. For example, GaAs and $Al_xGa_{1-x}As$ have different band gaps. What are the consequences for the motion of electrons (or holes) normal to the interface? Can charge carriers penetrate the interface?

The simplest theory visualized in Fig. 5.3 starts from the electron affinities χ_A and χ_B of the two materials A and B. The electron affinity is by definition the maximum energy one gains by adding an electron at rest from a region far away from the crystal to the neutral undoped semiconductor. The energy of the electron at large distance is called the vacuum level. The maximum energy is gained if the electron is filled into the bottom of the conduction band. If we consider two different materials, the vacuum level is the same for both. At the interface between two materials, the relative position of their conduction band minima is therefore given by the difference of the electron affinities, i.e., there arises a step in the conduction band edge, called conduction band offset, of size $\Delta E_c = |\chi_A - \chi_B|$.

For example, in GaAs we have $\chi_{GaAs} = 4.07$ eV, and for $Al_{0.3}Ga_{0.7}As$ we find $\chi_{AlGaAs} = 3.74$ eV. The resulting conduction band offset is $\Delta E_c = 330$ meV. The band gap difference between these materials is $\Delta E_g = 370$ meV, and therefore there will be a valence band offset $\Delta E_v = 40$ meV.

A word of caution is due here. Although this theory based on the electron affinities gives results of the right order of magnitude in most

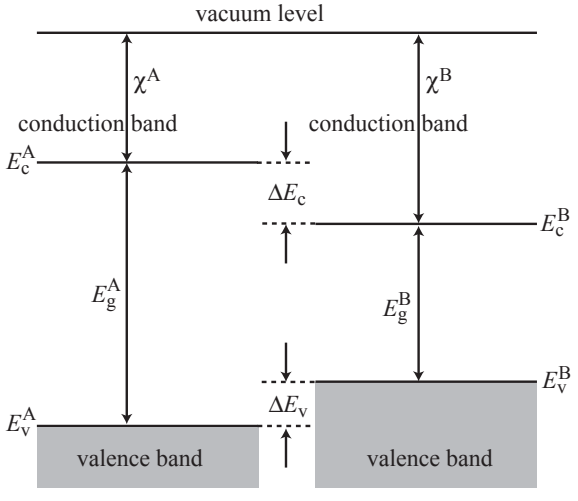


Fig. 5.3 Relative position of the band edges at a heterointerface (Davies, 1998).

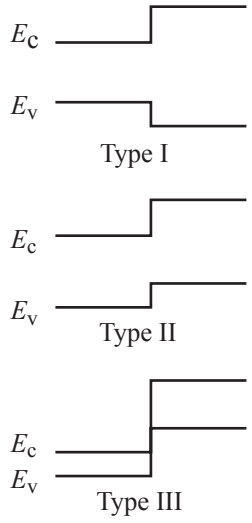


Fig. 5.4 Band line-up of conduction and valence band edges of two materials at a heterointerface of type I, II, and III.

cases, the band offsets that are determined experimentally can deviate significantly from this prediction. In the case of $GaAs/Al_{0.3}Ga_{0.7}As$, for example, $\Delta E_c/\Delta E_v = 0.62$ is the generally accepted value, while the simple theory gives $\Delta E_c/\Delta E_v = 0.85$.

In general, heterointerfaces belong to one of three categories (see Fig. 5.4). At a type I interface, such as in $GaAs/Al_{0.3}Ga_{0.7}As$, the conduction band edge is energetically higher, while the valence band edge is lower in one material than in the other. At an interface of type II, both band edges of one material are higher than in the other material; however, the valence band edge stays below the conduction band edge. If the latter condition is not fulfilled, we have a type III interface.

Envelope functions in the conduction band at heterointerfaces. Detailed theoretical investigations show, in agreement with experimental results, that the theory of envelope functions can also be applied at heterointerfaces if the conduction band minima (valence band maxima) of both materials are at Γ. This is not obvious because there is no translational symmetry of the crystal structure normal to the interface, the lattice periodic part of the wave functions can differ between the materials, and the perturbation given by the interface is typically not smooth on the scale of the lattice constant. Furthermore, as a result of the different band structures, the effective masses of the two materials will differ.

It turns out that the problem can be treated using the envelope function approximation if the boundary conditions of the envelope function at the interface are given by

$$F_c^{(A)}(\mathbf{r}) = F_c^{(B)}(\mathbf{r}), \tag{5.2}$$

i.e., the envelope function is steady at the interface, and

$$\begin{aligned} \mathbf{j}_A(\mathbf{r}) &= \tfrac{i\hbar}{2m_A^\star}\left(F_c^{(A)}(\mathbf{r})\nabla F_c^{(A)\star}(\mathbf{r}) - \text{ c.c.}\right) \\ &= \tfrac{i\hbar}{2m_B^\star}\left(F_c^{(B)}(\mathbf{r})\nabla F_c^{(B)\star}(\mathbf{r}) - \text{ c.c.}\right) \quad = \mathbf{j}_B(\mathbf{r}) \end{aligned} \tag{5.3}$$

for $\mathbf{r}$ at the interface, which guarantees that the probability current density is steady at the interface. Using eq. (5.2) in eq. (5.3) gives the condition

$$\frac{1}{m_A^\star}\nabla F_c^{(A)}(\mathbf{r}) = \frac{1}{m_B^\star}\nabla F_c^{(B)}(\mathbf{r}) \tag{5.4}$$

that has to be obeyed at the interface.

Quantum wells. In order to demonstrate the consequences of eqs (5.2) and (5.4) we consider a 10 nm GaAs layer (material B) sandwiched between thick $Al_{0.3}Ga_{0.7}As$ layers (material A). The band edge in the growth direction schematically shown in Fig. 5.5 has the shape of a potential well. Schrödinger's equation for the envelope function $F(\mathbf{r})$ in

the two materials reads

$$\begin{aligned} -\frac{\hbar^2}{2m^\star_{\rm A}}\Delta F(\mathbf{r}) &= (E - E_{\rm c}^{(\rm A)})F(\mathbf{r}) \text{ for } |z| > W/2 \\ -\frac{\hbar^2}{2m^\star_{\rm B}}\Delta F(\mathbf{r}) &= (E - E_{\rm c}^{(\rm B)})F(\mathbf{r}) \text{ for } |z| < W/2. \end{aligned}$$

The problem can be separated into three independent problems, one for each spatial direction. The envelope function can therefore be written as $F(\mathbf{r}) = \xi(x)\eta(y)\chi(z)$. As a result of translational invariance in x- and y-directions $\xi(x) = e^{ik_x x}$ and $\eta(y) = e^{ik_y y}$. The equation of motion in the z-direction is then

$$\begin{aligned} \frac{\hbar^2}{2m^\star_{\rm A}}\left[-\frac{\partial^2}{\partial z^2} + k_\parallel^2\right]\chi(z) &= (E - E_{\rm c}^{(\rm A)})\chi(z) \text{ for } |z| > W/2 \\ \frac{\hbar^2}{2m^\star_{\rm B}}\left[-\frac{\partial^2}{\partial z^2} + k_\parallel^2\right]\chi(z) &= (E - E_{\rm c}^{(\rm B)})\chi(z) \text{ for } |z| < W/2, \end{aligned}$$

where we have introduced $k_\parallel = \sqrt{k_x^2 + k_y^2}$. We are interested in bound states with $E_{\rm c}^{(\rm A)} > E > E_{\rm c}^{(\rm B)}$. The problem is symmetric with respect to $z = 0$. As a consequence, we expect wave functions with either even or odd parity, i.e., we use, in analogy with the standard quantum well problem,

$$\chi(z) = B \cdot \left\{ \begin{array}{l} \sin(k_z z) \\ \cos(k_z z) \end{array} \right. \quad \text{for } |z| < W/2$$

and

$$\chi(z) = Ae^{-\kappa|z|} \text{ for } |z| > W/2$$

with

$$\begin{aligned} k_z &= \sqrt{\frac{2m^\star_{\rm B}(E - E_{\rm c}^{(\rm B)})}{\hbar^2} - k_\parallel^2} \\ \kappa &= \sqrt{\frac{2m^\star_{\rm A}(E_{\rm c}^{(\rm A)} - E)}{\hbar^2} + k_\parallel^2}. \end{aligned}$$

The boundary conditions (5.2) and (5.4) at the interface require

$$\begin{aligned} Ae^{-\kappa W/2} &= B \left\{ \begin{array}{l} -\sin(k_z W/2) \\ \cos(k_z W/2) \end{array} \right. \\ \frac{A\kappa}{m^\star_{\rm A}} e^{-\kappa W/2} &= \frac{Bk_z}{m^\star_{\rm B}} \left\{ \begin{array}{l} \cos(k_z W/2) \\ \sin(k_z W/2) \end{array} \right. , \end{aligned}$$

which leads to the transcendental equation

$$\left\{ \begin{array}{l} -\tan(k_z W/2) \\ \cot(k_z W/2) \end{array} \right\} = \frac{m^\star_{\rm A}}{m^\star_{\rm B}} \frac{k_z}{\kappa}.$$

This equation can be solved numerically. The solution is given by energy eigenvalues of the form

$$E = E_n(k_\parallel) + \frac{\hbar^2 k_\parallel^2}{2m^\star_{\rm B}}$$

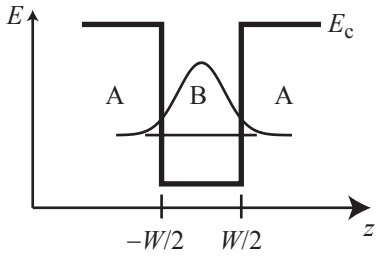


Fig. 5.5 Conduction band edge of an ABA type I heterostructure forming a quantum well, and the corresponding lowest bound state with its envelope function.

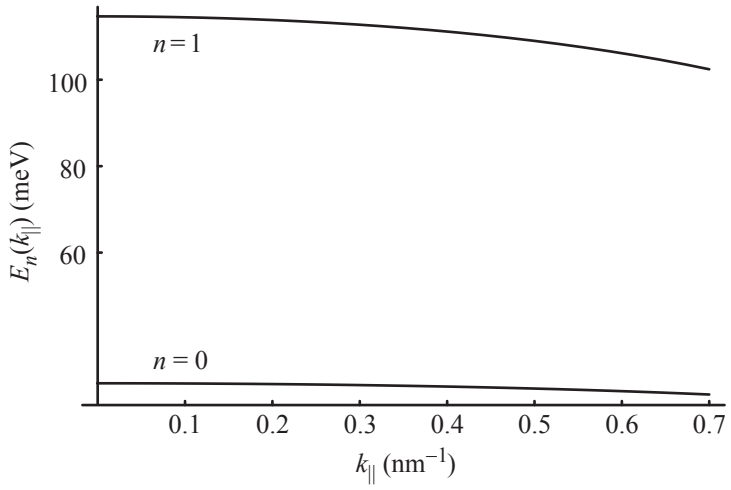


Fig. 5.6 Energy $E_n(k_\parallel)$ for the two lowest bound states plotted as a function of $k_\parallel = \sqrt{k_x^2 + k_y^2}$.

The distinct dispersion relations labeled by the quantum number n are called dispersion relations of *subband* n. The corresponding wave functions $\chi_n(z)$ are called *subband states* and the $E_n(k_\parallel = 0)$ are the *subband energies*. The dispersion $E_n(k_\parallel)$ is plotted in Fig. 5.6 using the material parameters of the Ga[Al]As material system. It is approximately parabolic, with negative curvature. The positive curvature of the total dispersion is therefore weakened, which can be interpreted as an increase of the in-plane effective mass. The electron acquires a larger effective mass because the wave function penetrates into the AlGaAs barrier where electrons have a larger effective mass than in GaAs. The mass increase for the lowest state is, however, quite small because of the exponential decrease of the wave function in the barrier. The total dispersion for the lowest three bound states is schematically plotted in Fig. 5.7. Two-dimensional electron systems in which only the energetically lowest subband is occupied with electrons are said to be in the *quantum limit*.

Simplified effective mass Schrödinger equation for heterostructures. As in the example discussed above, corrections to the in-plane effective mass in heterostructures are often quite small. It can be shown in a perturbative treatment that the in-plane effective mass $m_\parallel^\star$ is given by

$$\frac{1}{m_\parallel^\star} = \frac{p_\mathrm{A}}{m_\mathrm{A}^\star} + \frac{p_\mathrm{B}}{m_\mathrm{B}^\star}, \tag{5.5}$$

where p_A and p_B are the probabilities of finding the electron in material A or B, respectively, and $p_\mathrm{A} + p_\mathrm{B} = 1$. In cases where $p_\mathrm{A} \ll p_\mathrm{B}$ due to the small weight of the wave function in the barrier material B we find $m_\parallel^\star \approx m_\mathrm{B}^\star$. Similar considerations apply if one considers the problem of different effective $g^\star$ in a heterostructure.

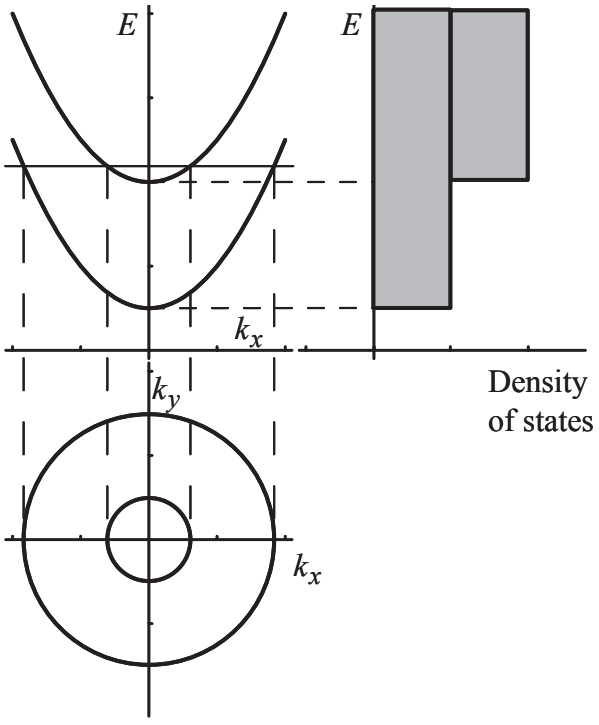


Fig. 5.7 Top left: Dispersion relations for the lowest two bound states in a quantum well heterostructure. Top right: density of states as a function of energy. Lower left: density of states as a function of energy.

Therefore, in many cases the problem of finding the bound states and energy levels in a heterostructure can be solved with good accuracy using the constant effective mass and $g^\star$ of the material in which the major part of the wave function resides. The simplified Schrödinger equation for electrons in a heterostructure with a spherically symmetric conduction band is then, in analogy with eq. (4.5), given by

$$\left[\frac{(\mathbf{p}+|e|\mathbf{A}(\mathbf{r}))^2}{2m^\star}+U(\mathbf{r})+E_{\rm c}(\mathbf{r})+\frac{1}{2}g^\star\mu_{\rm B}\sigma\mathbf{B}(\mathbf{r})\right]F(\mathbf{r})=EF(\mathbf{r}),\quad(5.6)$$

where $E_{\rm c}(\mathbf{r})$ is an effective potential arising from the spatially varying conduction band edge, and the material parameters $m^\star$ and $g^\star$ are given by the material where the dominant weight of the wave function is located.

Constant energy surfaces in heterostructures. In the quantum well example considered above, the motion of electrons in the growth direction (z-direction) is quantized and the motion in the plane of the quantum well is free. Therefore, the system is called two-dimensional. Constant energy surfaces in the k_x-k_y plane are circles, because the dispersion relation is isotropic in $k_\parallel$ (see Fig. 5.7).

Density of states for parabolic dispersions in two dimensions. The density of states for electrons in the quantum well heterostructure differs significantly from that found for three-dimensional systems [eq. (3.23)], or two-dimensional graphene [eq. (3.27)]. Assuming a two-dimensional spin-degenerate parabolic dispersion with in-plane effective mass $m^\star$

$$E_n(k_\parallel) = E_n + \frac{\hbar^2 k_\parallel^2}{2m^\star}$$

the integrated density of states below an energy E is given by

$$\begin{aligned} \mathcal{N}(E) &= \frac{2}{A} \sum_{\mathbf{k}, E_n(\mathbf{k})<E} 1 = \frac{2}{(2\pi)^2} \int d^2k \\ &= \frac{m^\star}{\pi\hbar^2} \int_0^E dE' = \frac{m^\star}{\pi\hbar^2} E. \end{aligned}$$

The resulting density of states for two-dimensional systems with parabolic dispersion is therefore

$$\mathcal{D}(E) = \frac{d\mathcal{N}(E)}{dE} = \frac{m^\star}{\pi\hbar^2}. \tag{5.7}$$

In contrast to the three-dimensional case in eq. (3.23), the density of states in two dimensions is independent of energy. If we take several subbands into account, the total density of states shows a number of steps. Figure 5.7 shows the density of states for a quantum well with more than one bound state.

Theoretical models with Bloch functions. More elaborate calculations of the band structure in heterostructures take the full Bloch functions into account. One can show that the concept of an envelope function still makes sense. As an example, Fig. 5.8 shows the result of such a calculation. The strongly oscillating part of the wave function is caused by the lattice-periodic part of the Bloch states. The envelope function does indeed show the kink in agreement with the boundary condition in eq. (5.4).

Two-dimensional hole gases. An effective confinement of holes in valence band quantum wells is also possible. Examples are heterostructures consisting of GaAs/AlGaAs, or Si/SiGe in which the lattice in the SiGe quantum well is strained due to lattice mismatch. The description of two-dimensional holes in an envelope function approximation is more complicated than that of electrons in the conduction band due to the four-fold degeneracy of states $|3/2, +3/2\rangle$, $|3/2, +1/2\rangle$, $|3/2, -1/2\rangle$ and $|3/2, -3/2\rangle$ at the top of the valence band. Calculations show that the quantization in the z-direction (normal to the heterointerfaces) lifts the degeneracy of light and heavy holes such that two pairs of degenerate states emerge. One of the degenerate pairs is at Γ made of heavy hole states $|3/2, +3/2\rangle$ and $|3/2, -3/2\rangle$. It is more strongly bound than the other pair of states which consists at Γ of light hole states $|3/2, +1/2\rangle$

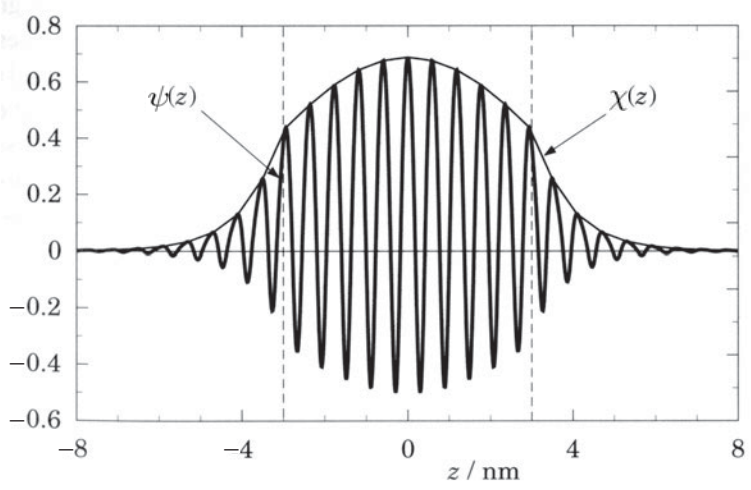


Fig. 5.8 Wave function of the lowest bound state in a 6 nm quantum well. The thin curve is the envelope function. (Reprinted with permission from Burt, 1994; copyright 1994, American Institute of Physics.)

and $|3/2, -1/2\rangle$. Most interestingly, the in-plane dispersion relation for the heavy and light hole quantized states shows *mass inversion*, i.e., the heavy hole subband has a smaller in-plane mass than the light hole subband. As a result, the two subband dispersions cross at finite wave vectors as shown in Fig. 5.9. At the crossing point, heavy and light hole states mix, and an avoided crossing results. In two-dimensional hole gases in the quantum limit, only the heavy hole subband is occupied.

Parabolic quantum wells. The material combination of GaAs and $Al_xGa_{1-x}As$ can also be used to create smooth potentials $E_c(z)$ in eq. (5.6). To this end, the aluminium fraction x is varied during growth in a parabolic fashion. The result is a parabolic conduction band edge, i.e., a parabolic confinement potential, because for $x < 0.45$ the band edge of $Al_xGa_{1-x}As$ varies linearly with the aluminium fraction x [see eq. (5.1)]. Such structures realize the potential of a harmonic oscillator in the growth direction of the crystal. However, as in the above example of the hard wall quantum well, the variation of the effective mass $m^\star$

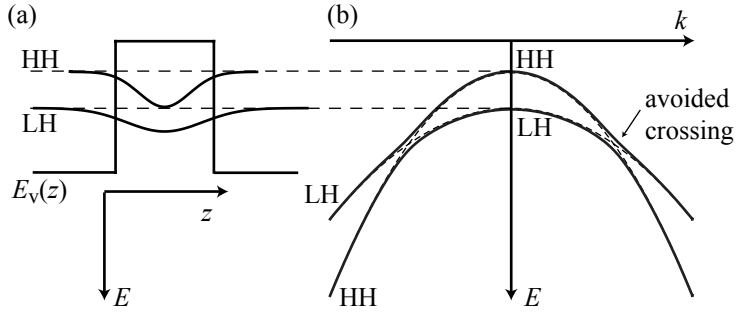


Fig. 5.9 (a) Subband energies and subband states of heavy (HHs) and light holes (LHs) in a quantum well. (b) Dispersion relations of heavy and light holes (Davies, 1998).

and $g^\star$ with aluminium fraction x along the growth direction makes the calculation of bound states more difficult. For small variations of x the mass variation can be neglected.

5.2 Doping, remote doping

Donors and acceptors. The properties of semiconductor materials can not only be changed during epitaxial growth by the choice and concentration of the material constituents, but also by a systematic incorporation of relatively small concentrations of doping atoms on specific lattice sites. Dopants that can release electrons into the conduction band through thermal activation (e.g., Si on a Ga lattice site in a GaAs crystal) are called donors. Typically these are atoms that possess one extra valence electron compared to the lattice atom that they replace. In this case we talk about n-doping. Dopants that thermally release holes into the valence band (e.g., B on a Si site in a Si crystal) are called acceptors. These are typically atoms that have one valence electron less than the lattice atom they replace. This type of doping is called p-doping. Many dopants form states in the band gap close to the bottom of the conduction band (shallow donors) or the top of the valence band (shallow acceptors). The states of shallow donors can often be described by the model of a hydrogen-like impurity (see page 57). Typically, shallow donors or acceptors are key ingredients for semiconductor nanostructures. Other defects form states that are deep within the band gap. These are called deep donors or deep acceptors. They typically reduce the material quality and are therefore undesired in semiconductor nanostructures. An example would be single missing crystal atoms (voids) that typically create states in the middle of the band gap.

Volume doping. If the dopants are evenly and statistically distributed in the crystal we talk about volume doping. It is characterized by the doping concentration which we will denote with N_D in the case of donors and N_A in the case of acceptors. The mean donor (acceptor) separation can be estimated by $d = N_\mathrm{D}^{-1/3}$ ($d = N_\mathrm{A}^{-1/3}$). If $d \gg a_\mathrm{B}^\star$ ($a_B^\star$ is the effective Bohr radius, i.e. the characteristic extent of the ground state wave function), the states of neighboring dopants typically do not overlap, and the system can be described as consisting of independent hydrogen-like impurities. As a consequence, the density of states has a sharp peak below the conduction band edge (in case of n-doping) at the energy of the hydrogen impurity ground state, as shown in Fig. 5.10(a). As the impurity concentration increases, the quantum states of neighboring doping atoms will start to overlap and the density of states peak becomes severely broadened as depicted in Fig. 5.10(b). If $d \ll a_B^\star$ the states of neighboring dopants do strongly overlap, their energy levels will split and the resulting density of impurity states below (for donors) or above (for acceptors) the band edges will be broadened even more, and electron (hole) states are no longer bound to individual dopants.

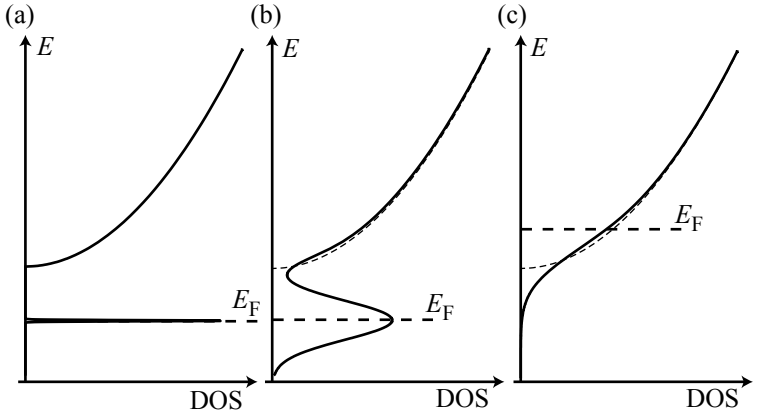


Fig. 5.10 Schematic plot of the density of states (DOS) near the conduction band edge for different doping levels in case of n-doping. (a) Low doping concentration. (b) Intermediate doping concentration. (c) Degenerate doping.

This latter case can be considered in the following way: the dopants create a spatially varying band edge energy. The spatially averaged contribution to this effective potential can be interpreted as a constant shift of the band edge into the band gap, or a smearing of the density of states as shown in Fig. 5.10(c). The electrons (holes) move in this potential as in the unperturbed conduction (valence) band. The remaining, spatially varying contribution leads to electron–dopant scattering, an effect that can often be treated by perturbation theory. As in a metal, the free electrons (or holes) will screen these potential fluctuations thereby reducing their influence on the electron motion. In this case, the Fermi energy at the temperature $T = 0$ lies above (below) the conduction (valence) band edge and we talk about degenerate electron (hole) gases, or degenerate doping.

In n-doped GaAs with $a_B^\star = 100$ Å we find $d \approx a_B^\star$ at a donor concentration of $N_D = 10^{18}$ cm^{-3}. This means that out of 64,000 lattice atoms one is a dopant and the relative concentration of donors is about 1.6×10^{-5}.

Sheet doping, δ-doping. During epitaxial MBE growth of a semiconductor crystal, dopants can also be incorporated in a plane. The doping profile in the growth direction of the crystal will then exhibit a sharp spike at the position of the doped plane. This way of doping a semiconductor is called δ-doping. As in the case of volume doping, the sheet doping concentration determines the material properties. The mean donor (acceptor) separation in the plane can now be estimated to be $d = N_D^{-1/2}$ ($d = N_A^{-1/2}$). If $d \gg a_B^\star$, the model of independent hydrogen-like impurities is appropriate. If $d \ll a_B^\star$, states of neighboring dopants overlap strongly. Again we can extract a mean potential created by the donors in the plane by spatial averaging (Jellium model). The remaining fluctuating part of the potential can again be treated as a perturbation of the free electron motion, which will be screened by the free electrons in the conduction band. In GaAs, the characteristic sheet doping concentration for donors is $N_D \approx a_B^{\star 2} = 10^{12}$ cm^{-2}.

Within the Jellium model the motion of electrons along an n-type doping plane can be regarded as that of free electrons with an effective mass $m^\star$. However, in the growth direction of the crystal, the electrons remain bound to the plane. This can easily be seen if we consider the electrostatic potential ϕ of a homogeneously charged sheet in a semiconductor. It is found as the solution of Poisson's equation

$$\frac{\partial^2\phi}{\partial z^2} = -\frac{|e|N_D\delta(z)}{\varepsilon\varepsilon_0}.$$

The solution that is symmetric around the doping plane is

$$\phi(z) = -\frac{|e|N_D}{2\varepsilon\varepsilon_0}\,|z|$$

and therefore the potential $U(z)$ for the electrons is

$$U(z) = -|e|\phi(z) = \frac{e^2N_D}{2\varepsilon\varepsilon_0}\,|z|\,. \tag{5.8}$$

This triangular potential binds the donor electrons to the doping plane as a heterostructure potential well created bound states binding electrons within the well.

Electron–electron interaction. The triangular potential in eq. (5.8) does not make physical sense because it becomes arbitrarily large for $|z| \to \infty$. The reason is that we have so far completely neglected effects of the electron–electron interaction. In particular, this is true for eq. (5.6), valid for electrons in the isotropic Γ-minimum of the conduction band. Including interactions on the approximation level of eq. (5.6) leads to a many-body effective mass hamiltonian for the envelope function.

$$\begin{aligned} H &= \sum_i \left[\frac{\left(\mathbf{p}_i + |e|\mathbf{A}(\mathbf{r}_i)\right)^2}{2m^\star} + U(\mathbf{r}_i) + E_{\mathrm{c}}(\mathbf{r}_i) + \frac{1}{2}g^\star\mu_{\mathrm{B}}\sigma_i\mathbf{B}(\mathbf{r}_i)\right] \\ &\quad + \sum_{\substack{i,\,j \\ i\neq j}} V_{\mathrm{C}}(\mathbf{r}_i - \mathbf{r}_j) \end{aligned} \tag{5.9}$$

In the simplest case the interaction between the electrons is described by the Coulomb interaction potential $V_{\mathrm{C}}(\mathbf{r}) = e^2/4\pi\epsilon\epsilon_0|r|$. A solution of the corresponding eigenvalue problem for the many-body envelope function cannot be found analytically and 'brute force' numerical solutions are restricted to small electron numbers. We will further discuss the general problem of solving Schrödinger's equation for interacting electrons, and the specific example of the delta doping layer in section 8.2.

Remote doping. The δ-doping technique has the big disadvantage that the motion of the electrons is bound to the doping plane in which the fluctuating part of the donor potential is strong. As a consequence, electrons are strongly scattered and the electrical resistivity of such layers

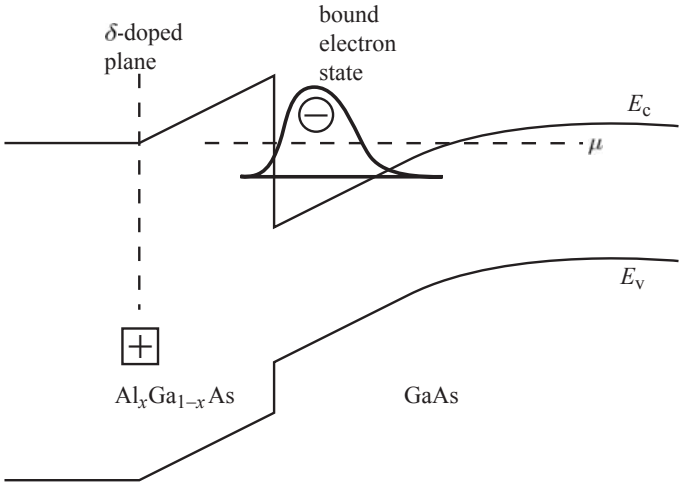


Fig. 5.11 Conduction band edge in growth direction of a heterostructure with remote doping.

is relatively high. Using a combination of δ-doping (or a thin volume doped layer) and a heterointerface, this disadvantage can be ruled out. This remote doping technique is shown in Fig. 5.11. It was introduced in 1978 by Dingle and Störmer (Dingle and Störmer, 1978). This doping technique implies that the doped layer is placed at a distance from a type I heterointerface. The undoped spacer layer between the heterointerface and the doping layer has a larger band gap than the material on the other side of the heterointerface. Due to the conduction (valence) band offset it is energetically favorable for the donor electrons (acceptor holes) to move to the material with the smaller band gap. The electrostatic attraction between positively charged donors (negatively charged acceptors) and electrons (holes), however, keeps the charge carriers close to the interface where, similar to the quantum well case, quantum confined states exist along the growth direction. In other words, there is an electric dipole formed by the plane with the charged donors and the two-dimensional electron gas.

As a result of the spatial separation between dopants and charge carriers, the fluctuating part of the dopant potentials in the plane of the charge carriers is strongly reduced from a divergent $1/r$ potential to a potential of the order of $e^2/(4\pi\varepsilon\varepsilon_0 d)$, where d is the separation between the two charged planes. Electron–dopant scattering is therefore also strongly reduced and the electrical resistivity of such a structure is considerably smaller than that of a pure δ-doped layer.

DX-centers in AlGaAs. So far we have mainly considered hydrogen-like shallow dopants. It turns out that Si-donors in $Al_xGa_{1-x}As$ also exist in a second state, the so-called DX-center. The crystal lattice deforms around the donor, if the latter gets occupied by an electron [see Fig. 5.12(b) and (c)]. This deformation is energetically more favorable, and the electron is more strongly bound than in a shallow donor. For example, the binding energy is $E_{DX} = 120\,\text{meV}$ for Si in $Al_{0.3}Ga_{0.7}As$.

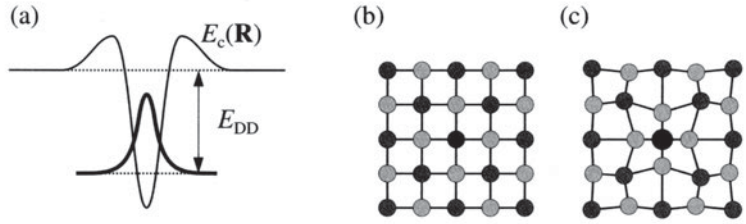


Fig. 5.12 (a) Schematic representation of the conduction band close to a DX-center. (b) Undistorted lattice for the case that the DX-state is not occupied. (c) Lattice deformation in case of the occupied DX-state (Davies, 1998).

The extent of the electronic donor state reduces to atomic dimensions. The binding energy of the DX-center depends on the aluminium fraction x. In pure GaAs the energy of the DX-center is above the conduction band edge, but it drops for $x > 0.2$ below the conduction band edge. Experimentally one finds that the occupation of DX-states freezes below about 150 K. This is due to the fact that, apart from the large binding energy, an activation energy is also required to occupy the state. Fig. 5.12 represents this situation schematically.

5.3 Semiconductor surfaces

Surface reconstruction. We now discuss the band structure near surfaces and its modifications compared to the bulk. As a first step we consider just the semiconductor surface without any metal evaporated on top. When the surface is formed, certain bonds remain without bond partners due to the termination of the crystal. These dangling bonds stick out of the surface and they would form energy bands of surface states resulting from their periodic arrangement. However, such bands are only called surface states if the resulting bands lie within the band gap of the bulk band structure, otherwise they are called surface resonances owing to their coupling to the states in the bulk. A surface can reduce its energy by rearranging the atoms at the surface, slightly allowing dangling bonds to mutually saturate each other. This process is called surface reconstruction. In this way, new unit cells form at the surface and the surface states are shifted in energy.

Electronic depletion and Fermi level pinning. In the case where no surface states exist within the band gap (this is, for example, the case on a clean freshly cleaved GaAs surface), the Fermi level in the bulk and at the surface are the same [see Fig. 5.13(a)]. If there are surface states, as is the case for most semiconductors, then the Fermi energies at the surface and in the bulk are not the same and charges are transferred between surface and bulk. As a result, electric fields arise and the bands bend. This situation is called Fermi level pinning in the band gap at the surface. The surface charge is neutralized by the charge of dopants in the volume of the crystal which results from the band bending.

It is of great importance for the understanding of semiconductor nano-

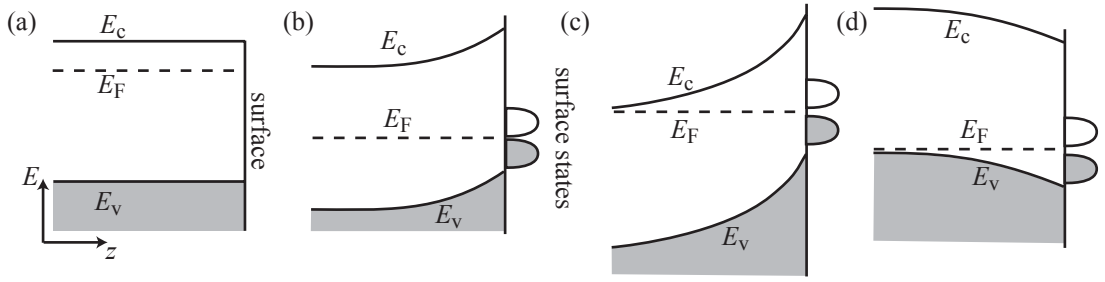


Fig. 5.13 (a) No surface states exist and therefore there is no pinning of the Fermi level at the surface. (b) Undoped semiconductor with surface states. The Fermi level in the bulk is in the middle of the gap. (c) The same as (b) for an n-doped semiconductor. (d) The same as (b) for a p-doped semiconductor (Yu and Cardona, 2001).

structures that in many cases the mere presence of a semiconductor surface depletes the underlying doped semiconductor material owing to the Fermi level pinning effect. This effect may still be present if oxides form at the surface or if metallic electrodes are evaporated onto the surface.

5.4 Metal electrodes on semiconductor surfaces

The deposition of metallic electrodes onto semiconductor surfaces is of great technological importance. In principle, there are two types of metallic electrodes: ohmic contacts and Schottky contacts which have rectifying properties.

Schottky contacts. Schottky contacts play an important role as gate electrodes in field effect transistors. A Schottky contact can be fabricated by evaporating a thin metal film onto a semiconductor surface. For example, a thin aluminium film can be evaporated onto a GaAs surface in a vacuum chamber with a background pressure of about 10^{-6} mbar.

It turns out that the energy difference between the Fermi level and the conduction band edge at the surface of some materials depends only weakly on the type of metal that is evaporated. The density of surface states is high enough to equilibrate the difference between the Fermi energies in the metal and at the semiconductor surface by charge transfer without a big energy shift of the Fermi level at the surface. For example, in GaAs a barrier between the Fermi energy in the metal and the conduction band edge of about 0.8 eV is formed almost independent of the metal. Figure 5.14 shows barrier heights for different metals on n-GaAs and n-Si.

We consider a simple model of a Schottky contact as depicted in Fig. 5.15. A metal contact is placed onto an n-doped GaAs substrate. The energetic separation from the Fermi level in the metal to the conduction band edge in the semiconductor is $\Phi_b = 0.8\,\text{eV}$. Far into the bulk of the semiconductor, the Fermi level is close to the conduction

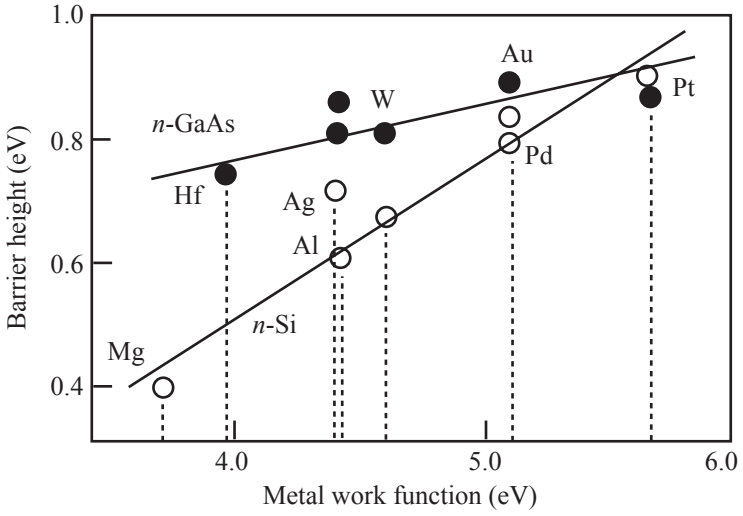


Fig. 5.14 Schottky barriers between n-GaAs or n-Si, and various metals, after Sze 1981.

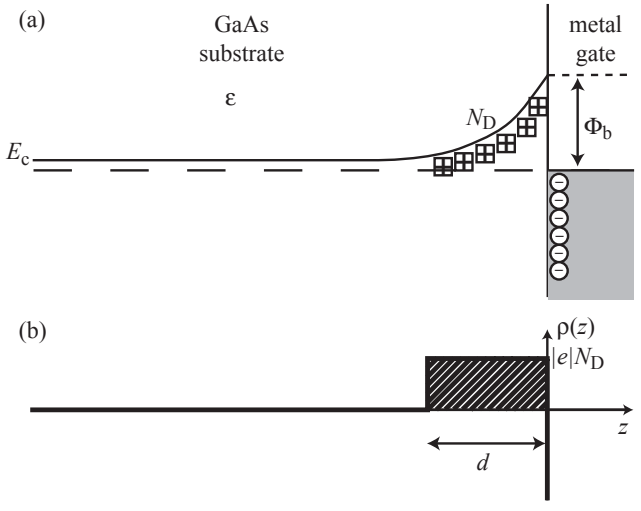


Fig. 5.15 Simple model of a Schottky contact between semiconductor and metal. (a) Conduction band profile and fixed charges. (b) Model charge distribution along z.

band edge, because of the n-doping. In thermodynamic equilibrium, the Fermi levels in the metal and in the semiconductor have to be the same. This is achieved by the presence of band bending in the semiconductor near the surface. It arises as a result of donor ionization in the vicinity of the surface leading to a space charge layer (depletion layer). In our simple model we assume that the space charge layer has a constant charge density $\rho_0 = eN_\mathrm{D}$ and a thickness d. According to Poisson's equation the resulting electrostatic potential near the semiconductor surface is parabolic with a maximum at $z = -d$:

$$\frac{\partial^2\phi(z)}{\partial z^2} = -\frac{|e|N_\mathrm{D}}{\varepsilon\varepsilon_0} \Rightarrow \phi(z) = -\frac{|e|N_\mathrm{D}}{2\varepsilon\varepsilon_0}\,|z+d|^2\,.$$

The effective potential seen by electrons in the conduction band is therefore $U(z) = -|e|\phi(z)$. The thickness d of the space charge layer adjusts in such a way that the potential at the surface at $z = 0$ obeys $-|e|\phi(0) = \Phi_\mathrm{b}$. This leads to

$$d = \sqrt{\frac{2\varepsilon\varepsilon_0\Phi_\mathrm{b}}{e^2N_\mathrm{D}}},$$

i.e., the larger the donor concentration, the smaller the thickness d of the space charge layer. If an additional voltage V_G is applied between metal and semiconductor, the boundary condition becomes $-|e|\phi(0) = \Phi_\mathrm{b} - |e|V_\mathrm{G}$ and the thickness of the depletion layer is

$$d(V_\mathrm{G}) = \sqrt{\frac{2\varepsilon\varepsilon_0(\Phi_\mathrm{b} - |e|V_\mathrm{G})}{e^2N_\mathrm{D}}}.$$

A negative voltage on the metal electrode increases the thickness of the depletion layer. This depletion of a semiconductor resulting from the voltage applied between semiconductor and metallic gate is called *field effect*. Assuming, for example, a doping concentration of $N_\mathrm{D} = 3\times10^{17}\,\mathrm{cm}^{-3}$ in GaAs we find at $V_\mathrm{G} = 0$ a depletion region with a thickness of $d = 61\,\mathrm{nm}$.

Field effect in a remotely doped heterostructure with a gate. Employing the field effect, the Schottky contact allows us to tune the electron density in a heterostructure with remote doping. The principle is illustrated in Fig. 5.16. The structure essentially operates like a parallel plate capacitor, where the metallic top gate and the two-dimensional electron gas bound to the heterointerface are the two capacitor plates. Applying negative gate voltages to the top gate the electron density can be reduced down to zero (complete depletion). With positive gate voltages the electron density can be increased until the electrons start to occupy the doped layer in the barrier. The tunability of the electron density is possible due to the Fermi level pinning. The Schottky barrier acts as an insulating barrier blocking electron transfer between the metal and the two-dimensional electron gas.

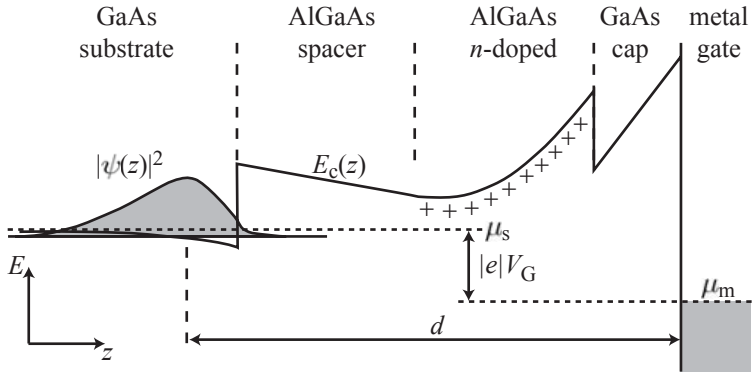


Fig. 5.16 Result of a self-consistent calculation of the conduction band edge in a heterostructure with remote doping and Schottky contact at the surface. The simplest way to describe the field effect in this structure is the parallel plate capacitor model. (Modelling program courtesy of G.L. Snider, University of Notre Dame.)

The simple plate capacitor model allows us to describe the tunability of the electron density as a function of gate voltage V_G. The capacitance per unit area of the barrier is given by

$$C = \frac{\varepsilon\varepsilon_0}{d},$$

where d is the separation of the metallized surface from the electron gas (we have assumed for simplicity that all materials forming the barrier have the same relative dielectric constant). The electron sheet density n_s is then given over a certain gate voltage range by the linear relation

$$n_s = n_s^{(0)} - \frac{\varepsilon\varepsilon_0}{ed} V_G. \tag{5.10}$$

A more complete description of the capacitive action of the top gate requires a quantum mechanical model. Such a model for the heterostructure will be extensively discussed in chapter 9.

Figure 5.17 shows the result of a measurement of the electron density in a remotely doped quantum well as a function of the applied gate voltage. The linear dependence of the density on gate voltage can be seen. As an example, we find with $\epsilon = 11.75$ for $Al_{0.3}Ga_{0.7}As$ and a barrier thickness of 40 nm a tunability of $1.6 \times 10^{12}\,\mathrm{cm}^{-2}$ per volt. Assuming a density $n_s^{(0)} = 5 \times 10^{11}\,\mathrm{cm}^{-2}$, about 300 mV applied voltage is sufficient to completely deplete the electron gas.

Electrons can penetrate the barrier region between metal and electron gas either by thermionic emission or by quantum tunneling. The latter process requires a sufficiently thin barrier, the former sufficiently high temperatures. At room temperature ($k_B T \approx 25\,\mathrm{meV}$), leakage currents due to thermionic emission are, for barrier heights of the order of 1 eV, relatively small. Quantum tunneling can be suppressed by growing sufficiently thick barriers.

Ohmic contacts. As a result of global charge neutrality, the depletion layer of a Schottky barrier becomes thinner, the larger the volume doping of the underlying semiconductor material is. If the doping concentration

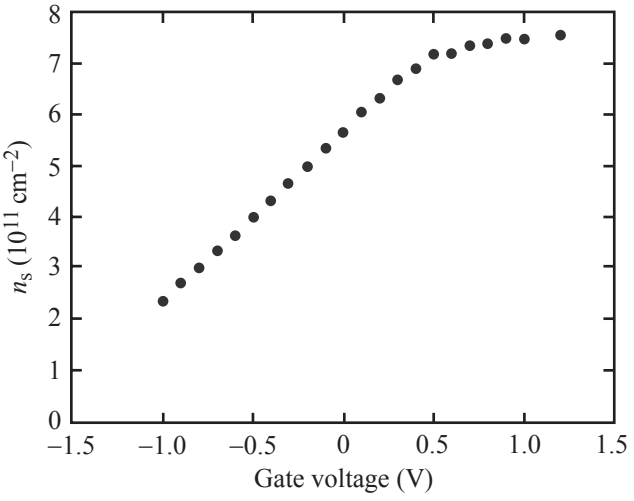


Fig. 5.17 Electron density in a 10 nm wide GaAs quantum well as a function of the gate voltage. The sheet density n_s was determined from the Hall effect.

is high enough, the tunnel barrier becomes sufficiently thin such that the Schottky contact no longer acts as an insulator and ohmic transport characteristics are found. This means that the current measured through the Schottky contact is proportional to the voltage.

One technological possibility for making an electrical contact to a semiconductor layer that has not been doped during growth is doping by alloying a metal. As an example, we consider a GaAs heterostructure (see Fig. 5.16) with remote doping. A so-called eutectic mixture of gold and germanium (this is an AuGe alloy with weight fractions 88% Au and 12% Ge) is evaporated onto the semiconductor surface. Then the sample is heated in an oven to about 450°C. Due to the high temperature, the germanium diffuses into the GaAs material and leads to n-doping with a concentration of $10^{19}\,\mathrm{cm}^{-3}$. Although the detailed mechanisms of this process remain to be investigated, it is believed that Ga leaves the crystal and diffuses into the gold layer, while the germanium atoms end up at Ga lattice sites and therefore act as donors. Depending on the duration and the exact temperature of this alloying process the ohmic contacts diffuse 100 nm–1 µm into the semiconductor. In this way, contact to buried two-dimensional electron gases can be made.

The quality of an ohmic contact is characterized by the contact resistance R_c. Obviously it is inversely proportional to the area A of the contact. The specific contact resistance r_c is therefore defined as

$$r_c = R_c A$$

and has the units $\Omega\mathrm{cm}^2$. Typical values for r_c in $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}/\mathrm{GaAs}$ heterostructures are $10^0 - 10^{-6}\,\Omega\mathrm{cm}^2$, depending on the doping level in the $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ barrier.

Further reading

- Band engineering: Singleton 2001; Weisbuch and Vinter 1991; Davies 1998; Heinzel 2007.
- Band offsets, donor levels, remote doping: Levinshtein *et al.* 1996.
- Surface states: Yu and Cardona 2001; Heinzel 2007.
- δ-doping: Schubert 1996.
- Ohmic contacts and Schottky contacts: (Williams 1990).

Exercises

(5.1) Consider the heterostructure potential well problem depicted in Fig. 5.5. Derive eq. (5.5) by treating the problem with different masses in the well and the barrier in first order perturbation theory. Use $1/m^\star_\mathrm{A} - 1/m^\star_\mathrm{B}$ as the small parameter.

(5.2) In this problem you will calculate the effective potential for electrons $U(z) = -e\phi(z)$ around a δ-doped layer at temperature $T = 0$ using the so-called Thomas–Fermi approximation. The δ-doped layer can be considered as a homogeneous sheet charge density $N_\mathrm{D}\delta(z)$ and the electronic charges bound to the doped layer are described by the electron density $n(z)$. Solve the problem by following these steps:

(a) The electrostatic potential $\phi(z)$ can be found from Poisson's equation. Write down this equation.

(b) In the Thomas–Fermi approximation, the electron density distribution $n(z)$ is found by integrating at each position z the three-dimensional density of states from the local band edge up to the Fermi energy, i.e.,

$$n(z) = \int_0^{E_\mathrm{F}+e\phi(z)} dE\mathcal{D}_\mathrm{3D}(E).$$

Solve this integral and insert the result in Poisson's equation. Simplify the resulting equation by substituting $e\tilde{\phi}(z) := E_\mathrm{F} + e\phi(z)$.

(c) Solve the resulting differential equation by integrating twice. Determine the integration constants by introducing physically motivated boundary conditions for $z \to \infty$ and for $z \to 0$. Hints:

(i) Solve the equation for $z \neq 0$ and treat the δ-function using a suitable boundary condition for $z \to 0$.

(ii) Simplify the problem by considering the symmetry around $z = 0$.

(iii) Substitute $x := \tilde{\phi}(z)$ and $y(x) := \partial\tilde{\phi}(z)/\partial z$.

Fabrication of semiconductor nanostructures

6

There are numerous methods of fabricating semiconductor nanostructures. This chapter can therefore only give prominent examples and show some of the most common techniques used in nanostructure fabrication. We distinguish bottom-up approaches, in which atoms are assembled into nanoscale structures during a growth process, and top-down approaches in which the nanostructures are carved out of macroscopic crystalline structures or defined electrostatically using patterned metallic electrodes. Sometimes, particular structures are obtained by a clever combination of the two approaches.

6.1 Growth methods

Self-assembly. Self-assembling growth (sometimes also called self-organized growth) of nanostructures in an MBE chamber belongs to the so-called bottom-up approaches in which structures are assembled from their atomic constituents during growth. As an example, we consider the growth of InAs on a GaAs substrate. The lattice constants of the two materials differ by 7%, leading to a strongly strained InAs layer. At an appropriate growth temperature, when the second monolayer forms, the InAs material releases this strain by forming small islands of 5–20 nm in diameter and a few nanometers in height. These islands are called self-assembled quantum dots[1] (SAQDs). The island formation is the result of a competition between binding energies on surfaces, at edges,

[1]The term *quantum dot* was to the best of the author's knowledge coined in (Reed *et al.*, 1986) as an extrapolation from 1D confinement in quantum wells, via 2D confinement in quantum wires to complete 3D confinement in quantum dots. It was quickly adopted by the scientific community. However, research on the physics of single electrons in completely confined structures already had a long-standing history at that time. For example, Millikan observed the effects of single electron charging on the speed of falling oil droplets (Millikan, 1911), single-electron tunneling in solids was studied by Gorter (Gorter, 1951) and later by Giaever and Zeller (Giaever and Zeller, 1968), and by Lambe and Jaklevic (Lambe and Jaklevic, 1969). Kulik and Shekhter had developed a detailed transport theory (Kulik and Shekhter, 1975). However, in the mid 1980s there was rapid progress in nanofabrication technology and the control over device parameters was improved tremendously.

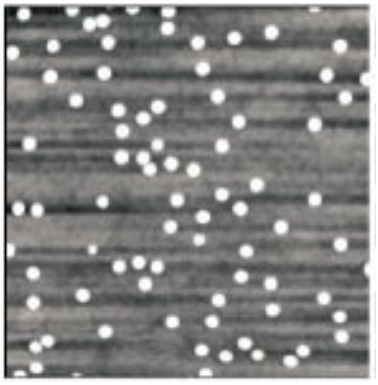


Fig. 6.1 AFM image of SAQDs on a GaAs (001) surface. The dots have a diameter of 30–40 nm, a height of 4–8 nm and they are statistically distributed on the surface. (Reprinted with permission from Petroff *et al.*, 2001. Copyright 2001, American Institute of Physics.)

and in the volume, as well as strain energies. The elastic energy of the strained two-dimensional InAs layer on GaAs grows with the square of the layer thickness. The total energy of the system can therefore be reduced by the formation of edges and islands at low thicknesses. Most interestingly, these self-assembled quantum dots have a perfect crystal structure without lattice defects. This regime of crystal growth where lattice mismatched material spontaneously forms islands is called the Stranski–Krastanov growth mode. If one continues the growth of the GaAs crystal after the formation of the SAQDs, they become embedded in a GaAs matrix with nanometer-sized InAs enclosures. Typically, the sheet density of these dots is in the range between 5×10^9 and $1 \times 10^{11}\,\mathrm{cm}^{-2}$.

Figure 6.1 shows an image of InAs quantum dots on a GaAs (100) substrate. The dots are statistically distributed on the surface. A regular arrangement of dots can be achieved by prepatterning the substrate (see below).

Electrons and holes can be trapped on the InAs islands due to the smaller band gap of InAs as compared to GaAs (type I heterostructure). An electron experiences a locally strongly reduced potential, i.e., potential well with confinement in all three spatial dimensions. Bound states form in this well similar to the situation in an atom.

The calculation of the electronic structure of SAQDs is complicated by the fact that the exact geometry and material composition are often not exactly known. For example, in InAs SAQDs on GaAs, gallium can easily get alloyed into the dot material and form the ternary semiconductor $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$. Additional complications arise because the material is strained, and piezoelectric effects enter. The results of calculations using the effective mass approximation have been found not to be reliable.

Experimentally, individual SAQDs could be investigated. In optical experiments, extremely sharp emission lines of individual dots were found in the photoluminescence, and they resemble atomic transitions. SAQDs have also been used as the active material in the resonators of semiconductor lasers. The transport properties of individual SAQDs were investigated by single-electron tunneling experiments.

The number of materials for which this growth technique can be exploited is large. SAQDs have been realized with Si/SiGe, III-V semiconductors and II-VI materials. The challenge for materials engineers is the growth of dots with a very narrow size and shape distribution and with homogeneous chemical composition. Another challenge is to predetermine the position of the SAQDs in the plane.

Prepatterned substrates. Nanostructures can also be fabricated by MBE-growth on prepatterned substrates. An example is the realization of quantum wires on a (100) GaAs substrate. Long V-shaped grooves oriented in the [011] direction are defined by electron beam lithography (see below) and subsequent etching. On this prepatterned substrate, a periodic sequence of AlGaAs/GaAs quantum wells is grown. Figure 6.2(a) shows a cross-sectional view of such a structure. One can see

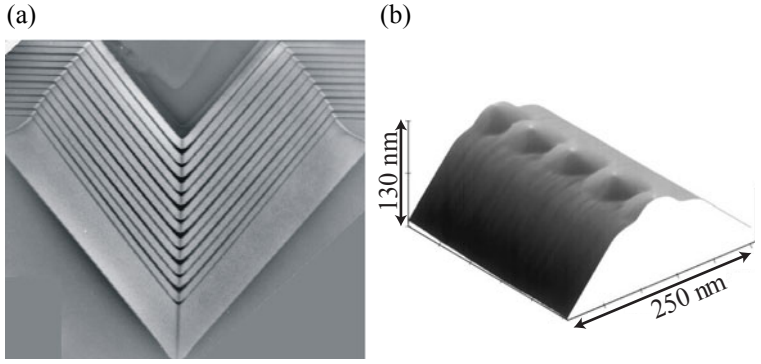


Fig. 6.2 (a) Image of GaAs quantum wires grown on a prepatterned substrate. The cross-sectional image was taken with a transmission electron microscope. Image courtesy of D. Meertens, Forschungszentrum Jülich. (b) AFM image of InAs quantum dots grown on a prepatterned InP substrate. (Reprinted from Williams *et al.*, 2001 with permission from Elsevier.)

that at the bottom of the V-groove the (dark) GaAs layers are thicker than at the tilted side walls. As a consequence, the electronic states at the bottom of the groove are energetically lowered, and one-dimensional bound states form along the groove.

Prepatterning a substrate before the self-assembled growth of quantum dots can lead to ordering of the SAQDs. Figure 6.2(b) shows an example where a well-defined ridge has formed on an InP substrate. If one grows InAs on this ridge under appropriate growth conditions, SAQDs form along the ridge with a quite regular mutual separation.

Cleaved-edge overgrowth. A technique called cleaved-edge overgrowth (CEO) was developed at the beginning of the 1990s at the Bell Labs. It allows the fabrication of very high quality quantum wires and quantum dots. The method starts with the MBE-growth of a conventional GaAs/AlGaAs quantum well. Subsequently the material is in-situ cleaved along the (110) direction (see Fig. 6.3). The growth continues on top of the cleaved surface with another quantum well. In this way, two orthogonal electronic systems of the highest quality meet along a line. Along this line one-dimensional bound states form.

If this crystal is cleaved again orthogonally to the previous cleavage, the growth can be continued with a third quantum well. The three resulting wells touch at a single point where bound zero-dimensional quantum dot states form (see Fig. 6.4). If two parallel quantum wells are grown, two quantum dots coupled by tunneling or by electrostatic interaction can be created.

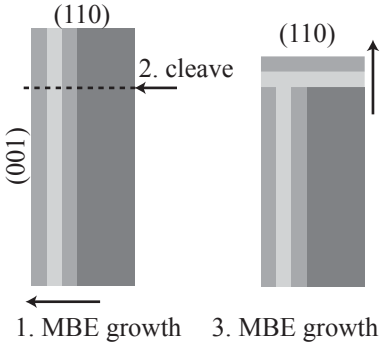


Fig. 6.3 Schematic presentation of the cleaved-edge overgrowth (CEO) method.

Catalytic growth of nanowires. Nanowires can also be grown using the so-called vapor–liquid–solid growth mode. Aerosol gold particles with size in the range between a few nanometers and a few tens of nanometers are deposited onto a crystalline III-V semiconductor substrate oriented along the (111) direction. This substrate is then transferred to the growth chamber of a MOVPE system (metal organic vapor phase epitaxy). When constituents of a semiconductor crystal are provided they will diffuse into the gold particles. Once a critical saturation

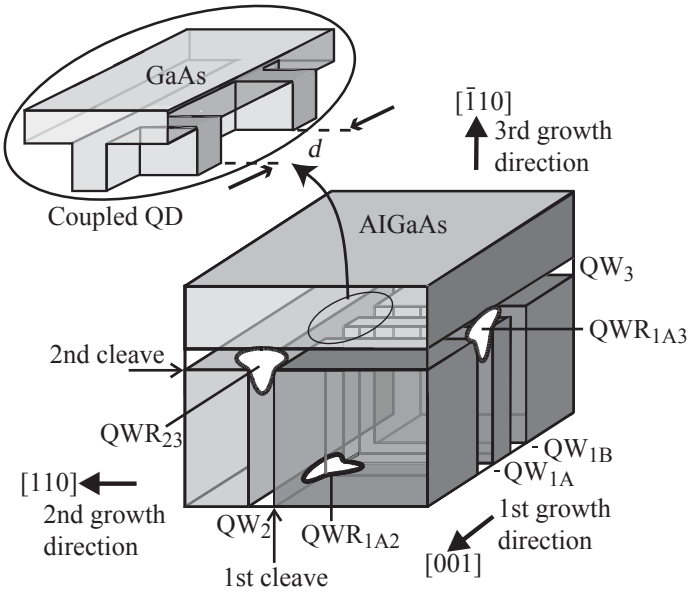


Fig. 6.4 Scheme of a complicated nanostructure fabricated by applying a two-step CEO process. At each crossing point of three orthogonal quantum wells, a quantum dot forms. In this structure, a coupled quantum dot system was realized (Schedelbeck *et al.*, 1997).

concentration is reached in the particle, crystal growth sets in at the interface between the substrate and the gold particle. In this way, high quality crystalline semiconducting nanowires can be grown normal to the substrate between the substrate and the gold particles. The diameter of these wires, which are also referred to as *nanowhiskers*, is determined by the size of the gold particle; their length is determined by the growth rate and the growth time. Lengths of several micrometers have been reported. While a random deposition of aerosol gold particles creates a random distribution of the nanowires, ordered arrays of gold catalyst can be created using electron beam lithography techniques (see below). A growth method has been reported in Mandl *et al.* 2006, where instead of the gold catalyst, a thin SiO_x-layer has been deposited on the substrates prior to the MOVPE wire growth. Nanowhiskers of many different materials have been studied, such as Si, Ge, GaAs, GaP, and InAs. A review of the growth and optical properties of GaAs and InAs whiskers can be found in Sato *et al.* 1995.

During nanowire growth the material composition can be changed, and abrupt heterointerfaces have been realized normal to the wire axis. Figure 6.5 shows a cross-section through an InAs nanowire with barriers made of InP. In this image the catalytic gold particle on top of the nanowire can be seen. Using this growth technique incorporating heterointerfaces, quantum dot structures and resonant tunneling structures can be made.

Such nanowires have been characterized with optical techniques and transport experiments. For the latter, the nanowires are typically broken off the substrate and put with the wire axis parallel to the surface of another substrate. The wires can then be located and contacted with

the lithographic techniques described below. In nanowires without heterointerfaces, quantum dots can be induced with laterally patterned gate electrodes. Possible applications of these nanowires are field effect transistors with particularly low power consumption, memory applications, light emitters, and nanosensing.

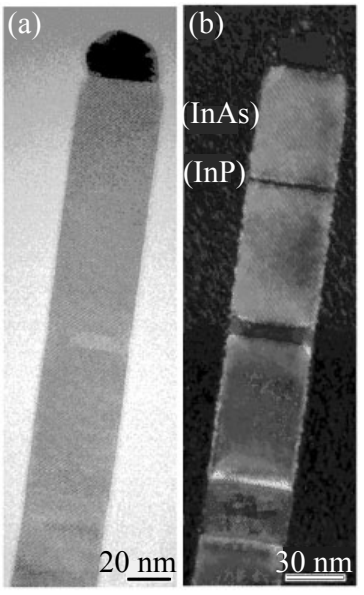


Fig. 6.5 (a) High resolution transmission electron microscope image of an InAs nanowire with several InP/InAs heterointerfaces. (b) Color-coded image (printed here in grayscale) in which the InP regions are emphasized. (Reprinted with permission from Bjork *et al.*, 2002. Copyright 2002 American Chemical Society.)

Carbon nanotube fabrication. Carbon nanotubes (CNTs) can be viewed as two-dimensional graphene sheets (i.e., monolayer graphite) that are rolled up to form seamless cylindrical tubes of diameters in the nanometer range and lengths of many micrometers up to a millimeter. They exhibit an enormous tensile strength and have interesting electronic and optical properties.[2] Two types of CNTs can be distinguished. Single-walled nanotubes (SWNTs) consist only of a single graphene sheet, whereas multiwalled nanotubes (MWNTs) are made of several coaxial single-wall tubes, the outer ones surrounding the inner ones.

Depending on the direction in which the graphene sheets are rolled up, SWNTs can be classified using the so-called chiral vector. If $\mathbf{a}_1$ and $\mathbf{a}_2$ are the two unit vectors of the honeycomb graphene lattice (see Fig. 3.5), a chiral vector can be represented as $\mathbf{C} = n\mathbf{a}_1 + m\mathbf{a}_2$, with n and m being integer. A CNT in which the start and end point of $\mathbf{C}$ coincide for certain (n, m) is called an (n, m)-nanotube. *Armchair nanotubes* have $n = m$ and *zig-zag nanotubes* $(n, 0)$. Other nanotubes are called *chiral*. Figure 6.6 shows a scanning tunneling microscope image of a chiral carbon nanotube with atomic resolution. Electronically, SWNTs are metallic for $2n + m = 3k$ (k integer) and semiconducting otherwise.

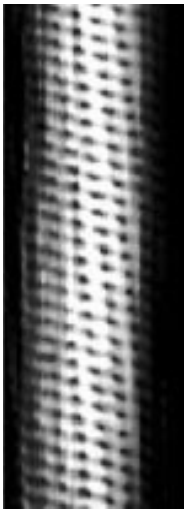


Fig. 6.6 Scanning tunneling microscope image of a CNT with atomic resolution. (Wildoer *et al.*, 1998. Reprinted by permission from Macmillan Publishers Ltd. Copyright 1998.)

CNTs can be produced in a variety of ways. Initially they were found to be produced in arc discharges using carbon electrodes. This method is probably most widely used. SWNTs and MWNTs are produced at the same time in a random mixture.

An alternative production method is laser ablation from a graphite target. A pulsed laser vaporizes the graphite in a high temperature reactor which is flushed with an inert gas. Nanotubes form at the cooler walls of the reactor chamber. This method has a yield of about 70% and produces predominantly SWNTs. The tube diameter can be controlled to some extent by the reaction temperature.

Chemical vapor deposition has also been used to produce CNTs. This growth method uses a substrate on which catalytic particles, for example, nickel, cobalt, or iron, are prepared. The size of these particles is related to the diameter of the CNT that will grow. Carbon is provided in the growth chamber as a carbon containing gas, such as ethanol, ethylene, or acetylene. At the sites of the catalyst particles, the gas molecules are broken apart and the nanotubes grow in random directions at the edges of the particle.

[2]For example, they were used to enhance the frame of the carbon fiber bicycle frame of F. Landis used in the 2006 Tour de France.

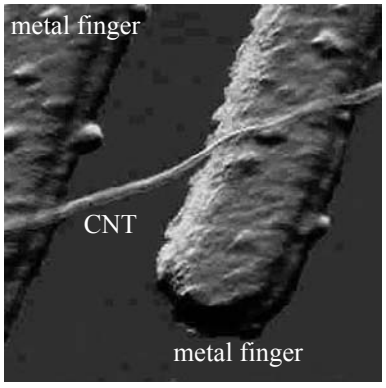


Fig. 6.7 Scanning force microscope image of a CNT connecting two platinum contact fingers (Tans *et al.*, 1997. Reprinted by permission from Macmillan Publishers Ltd. Copyright 1997.)

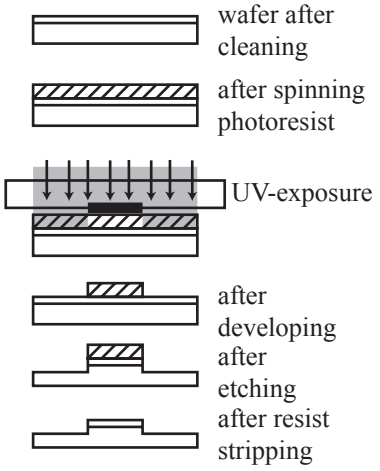


Fig. 6.8 Lateral patterning of a semiconductor wafer with photolithography and etching.

If strong electric fields are applied during the growth, a plasma can be created leading to the plasma enhanced chemical vapor deposition method. In this case the nanotubes grow in the direction of the electric field.

For making electrical contacts to carbon nanotubes, they are brought into solution and spread onto a suitable substrate. Individual tubes can then be located with a scanning electron microscope. Metallic contacts to the tubes are made with lateral patterning techniques and metal evaporation. Figure 6.7 shows an example of a CNT connecting two Pt electrodes on a Si/SiO_2 substrate.

6.2 Lateral patterning

Apart from the growth methods for the fabrication of nanostructures introduced in the previous section, there are very common and technologically mature methods for laterally structuring substrates. In industry, mainly photolithographic methods are used. In research, more flexible methods, such as electron beam lithography (EBL), or local anodic oxidation with a scanning force microscope (AFM lithography) are employed. From the processing perspective three techniques can be distinguished: we can either remove material (e.g., by wet chemical etching), deposit material (e.g., by evaporating metal electrodes), or modify the material locally (e.g., by local oxidation or by ion implantation).

Photolithography. Using the technique of photolithography, a pattern can be transferred from a so-called mask onto a thin layer of photoresist. The process is similar to the preparation of a paper copy of a photograph. The mask acts as the negative, the substrate with the photoresist corresponds to the photographic paper. The photoresist is exposed to UV-light through the mask and then developed. The developer (a suitable chemical agent) dissolves the resist in those regions where it was exposed and the semiconductor surface is uncovered locally. In this way the resist pattern becomes an image of the mask pattern.

Figure 6.8 shows the relevant steps of a photolithographic process with subsequent etching of the semiconductor:

(1) wafer cleaning
(2) spinning photoresist
(3) exposure of the resist through the mask
(4) developing the resist
(5) etching of the uncovered surfaces

One distinguishes positive resists, negative resists, and image reversal resists. The positive resist dissolves in the exposed regions during development (see Fig. 6.8), whereas negative resist dissolves in the regions that were not exposed.

A typical positive resist is made of a light-sensitive compound, a base material and a suitable organic solvent. The developer rapidly dissolves the base material in the presence of the light-sensitive compound (e.g., 15 nm/s), in its absence more than a hundred times worse (e.g., 0.1 nm/s). The resist is exposed with light of a wavelength between 300 and 450 nm, i.e., in the UV-range (mercury-vapor lamp). It destroys the light-sensitive compound in such a way that the exposed resist becomes soluble. The wavelength of the light limits the spatial resolution of the method through diffraction and interference effects. With shorter wavelengths and refined photoresists a resolution well below 100 nm is reached in industry today.

The sample is covered with the photoresist using a resist spinner. Rotation speeds between 2000 and 8000 rpm lead to resist thicknesses between 2.5 µm and 300 nm.

If metallic electrodes need to be deposited on a semiconductor wafer, the so-called lift-off technique is used. Figure 6.9 shows the essential steps of this process. Crucial for this process is that after the development step the photoresist edges have an undercut profile. If a sufficiently thin metal film is evaporated, it will not be continuous at the resist edge. Upon resist removal, the parts of the metal film on top of the resist will be lifted and removed with the resist.

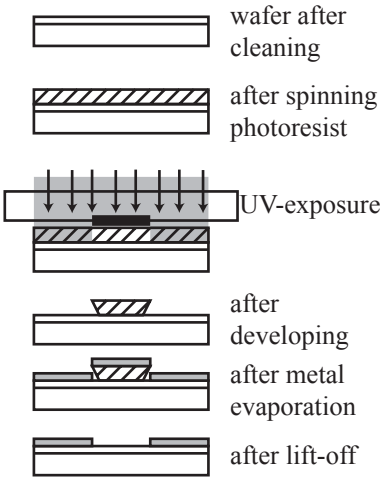


Fig. 6.9 Lateral patterning of gate electrodes with the lift-off technique.

Electron beam lithography (EBL). Special resists exist that are not made for exposure with light but with electrons of a certain energy (typically 10 to 25 keV). For the exposure, the beam of a scanning electron microscope can be used. The position of the beam can be exactly controlled and as a result one can directly write a pattern into the resist without the need of a mask. Therefore electron beam lithography is a very flexible method which is widely used in research. With the best microscopes, structures down to the size of about 30 nm can be written.

As in photolithography, positive and negative resists are used, made on the basis of polymers. The most common positive resist is PMMA (polymethyl methacrylate, or polymethyl-2-methylpropanoate, also known as acrylic). In positive resists, chemical bonds are cracked by the impinging electrons and the exposed region is more soluble. In negative resists, the exposure leads to a strong cross-linking of the molecules and, as a result, to a weaker solubility. The dose for an exposure is typically in the range of 2×10^{-7} to 8×10^{-7} C/cm^2.

Hall bar structure: As an example for the application of the techniques introduced above, we describe the fabrication of a so-called Hall bar structure. The following processing steps are necessary (see also Fig. 6.10):

(1) Mesa patterning

- wafer cleaning
- spinning photoresist

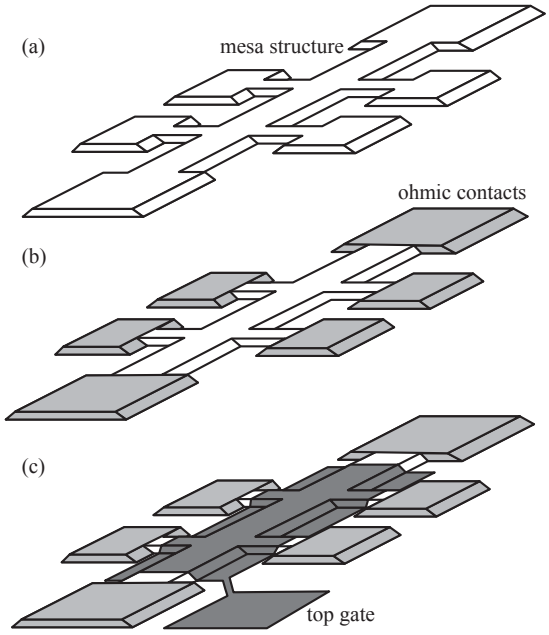


Fig. 6.10 Step by step fabrication of a Hall bar structure with top gate. (a) Mesa structure. (b) After the fabrication of ohmic contacts. (c) After evaporating the top gate.

- exposure of the resist through the mask with the mesa structure
- developing the resist
- etching of the uncovered semiconductor surfaces
- removing the remaining photoresist

(2) Ohmic contacts

- wafer cleaning
- spinning photoresist
- exposure of the resist through the mask with the ohmic contact structure
- developing the resist
- evaporating gold–germanium–nickel
- lift-off process
- alloying contacts

(3) Top gate

- wafer cleaning
- spinning photoresist
- exposure of the resist through the mask with the gate structure
- developing the resist
- evaporating titanium–platinum–gold
- lift-off process

Split-gate technique. Using the so-called split-gate technique, very narrow quantum channels for electrons can be fabricated on the basis of two-dimensional electron gases. To this end, two (or more) finger-shaped gate electrodes are evaporated in such a way that only a narrow channel remains in-between. An example is shown in Fig. 6.11. The two-dimensional electron gas below the gate can be completely depleted by applying negative voltages on the three gate electrodes. The electrons can only pass through the narrow channels between the gates. Such a narrow channel is called a quantum point contact. Figure 6.12 shows a cross-sectional view through such a channel.

Fabricating a number of split-gates allows us to define quantum dots. These structures are based on two-dimensional electron gases and six gate fingers are evaporated as shown in Fig. 6.13. The electron gas is depleted by applying negative voltages to the finger gates. Owing to the larger separation of the two middle fingers as compared to the two outer pairs of gates, electrons can be localized on an island between the gates. The two outer pairs of electrodes determine the coupling of the island to the extended electron gas outside the quantum dot.

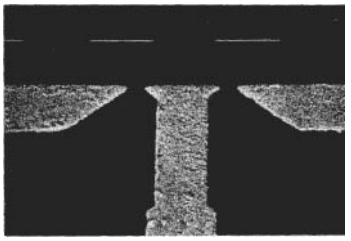


Fig. 6.11 Two quantum point contacts defined with the split-gate technique. In the narrowest region, the separation of the two electrodes is 250 nm. (Reprinted with permission from van Houten *et al.*, 1989. Copyright 1989 by the American Physical Society.)

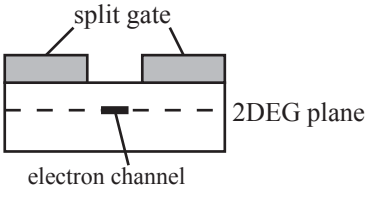


Fig. 6.12 Cross-sectional view of a heterostructure with a two-dimensional electron gas and a split-gate on the surface. Negative voltages on the gate electrodes deplete the electron gas and a narrow electronic channel forms.

AFM lithography. AFM lithography is a method for the fabrication of semiconductor nanostructures which is innovative, very flexible, and simple in principle. It is not used industrially, but a number of research labs have developed the relevant know-how for using the method routinely for nanostructure fabrication.

Figure 6.14 shows schematically how this technique works. The basis is, for example, a GaAs/AlGaAs heterostructure with a shallow two-dimensional electron gas (about 40 nm below the surface). At room temperature a water film will be present on the surface. Its thickness can be controlled via the humidity of the air.

The tip of a scanning force microscope is positioned close to the surface. Using the piezoelectric actuators of the microscope, the tip can be moved laterally along the surface of the substrate. A control loop measures the force between tip and surface and keeps the tip–sample separation constant during the lateral motion using another piezoelectric actuator (z-piezo). By scanning the tip line by line above the surface and measuring the voltage applied to the z-piezo, one obtains a map of the surface topography. During such mapping the forces between the tip and the surface are in the range of a few nanonewtons and no mechanical wear arises. Lateral resolutions in the nanometer range are routinely achieved, atomic resolution is possible. Vertically, the resolution can be increased into the subangstrom range.

If doped silicon tips or metal coated tips are used, voltages can be applied between the tip and the sample. An appropriate voltage (typically between -10 and -20 V) results in a local oxidation of the sample surface below the tip. For this process, the water film plays an important role. The piezoelectric actuators are used to move the tip slowly over the surface. In this way, oxide lines of arbitrary shape can be written on the surface. Oxide line widths of about 100 nm are routinely achieved.

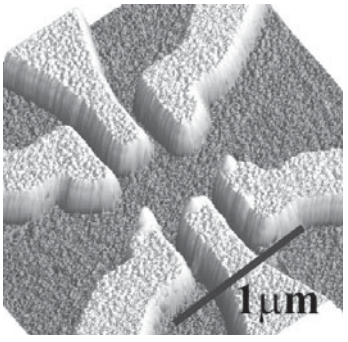


Fig. 6.13 GaAs/AlGaAs heterostructure with a two-dimensional electron gas below the surface and a split-gate defined quantum dot. Negative voltages applied to the gate electrodes deplete the underlying electron gas such that electrons are localized on the island between the two central gate fingers.

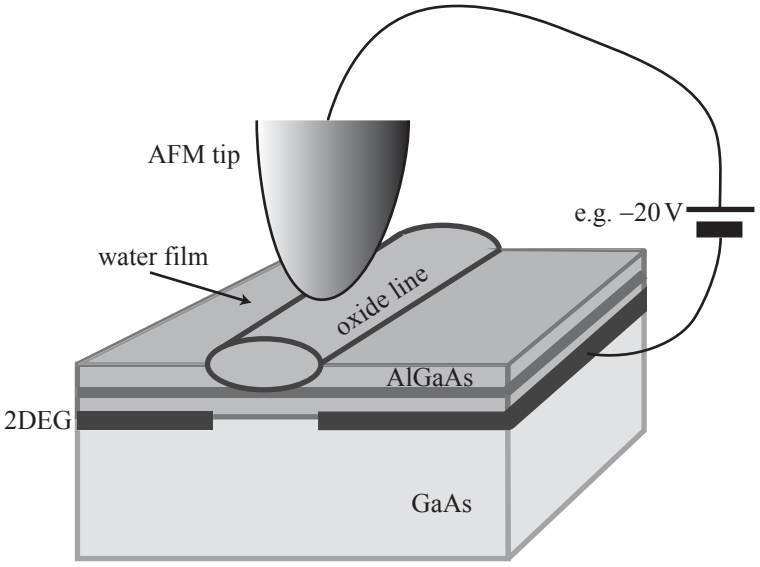


Fig. 6.14 Principle of AFM lithography.

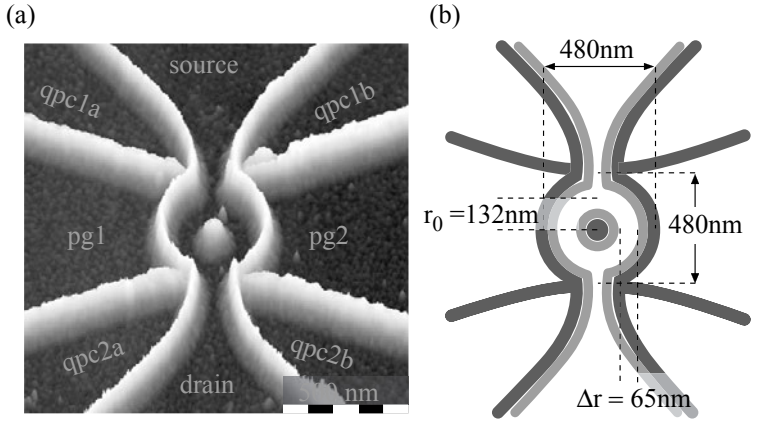


Fig. 6.15 (a) Quantum ring structure fabricated by AFM lithography on a GaAs/AlGaAs heterostructure. (b) Schematic picture of the structure where all the length scales are indicated (Fuhrer *et al.*, 2001).

The height of the lines can vary between 2 and 30 nm.

It turns out that the electron gas is completely depleted below the oxide lines. In this way, oxide lines split the electron gas into electrically separate parts. At liquid helium temperatures, a few hundred millivolts can be applied between these parts, before a measurable leakage current flows. By arranging the oxide lines on the surface in a suitable way, many types of nanostructures can be fabricated. Among them are quantum point contacts, quantum wires, quantum dots, and quantum rings. Figure 6.15 shows, as an example, a quantum ring fabricated in this way. The regions labeled 'qpc1a', 'qpc1b', 'qpc2a', 'qpc2b', 'pg1', and 'pg2' denote parts of the two-dimensional electron gas that can be used as in-plane gates for tuning the electron density in the structure. Between the source and the drain contact, a voltage can be applied for measuring the conductance of the ring.

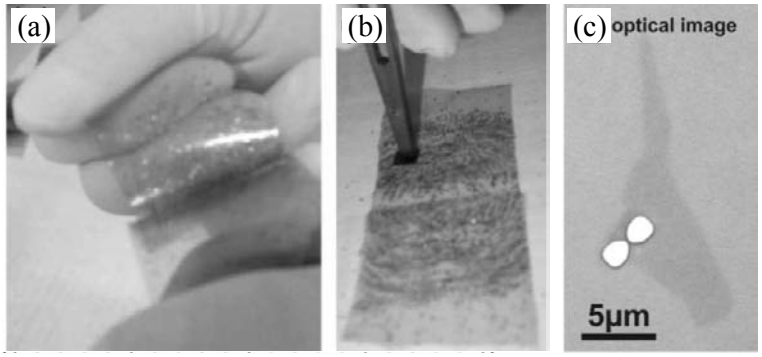


Fig. 6.16 (a) Folding and unfolding the adhesive tapes reduces the number of layers per flake. (b) Flakes are transferred onto a substrate by pressing it on the adhesive side of the tape. (c) Single layer flakes can be discerned under an optical microscope.

Exfoliation of graphene. A particularly simple technique is used to deposit single-layer graphene flakes on substrates. The starting material is, for example, a powder consisting of natural graphite flakes. The powder is distributed on an adhesive tape. Subsequent folding and unfolding the tape [see Fig. 6.16(a)] tears the stacked graphene sheets apart and therefore leads to thinning of the graphene flakes, some of which can be single layer, sitting on the sticky side of the tape (Novoselov *et al.*, 2004). This material is then transferred onto a highly doped silicon substrate which is covered by a 300 nm SiO_2 layer by pressing the oxidized surface onto the tape [Fig. 6.16(b)]. The thickness of the oxide layer on the substrate has been chosen such that a single-layer graphene flake on the surface can be distinguished from thicker flakes and from the bare surface under an optical microscope by using a green filter [see Fig. 6.16(c)]. A metal grid fabricated on the oxide by photolithography before the transfer of the flakes allows one to describe the position of a flake, once it has been found. Later on, the flake can be contacted and patterned by electron beam lithography and metal evaporation or etching.

Further reading

- Weisbuch and Vinter 1991; Heinzel 2007; Williams 1990.

This page intentionally left blank

Electrostatics of semiconductor nanostructures

7

7.1 The electrostatic problem

The determination of static electric fields, or the electric potentials within a nanostructure is a well-known problem of electrostatics. Figure 7.1 shows a typical arrangement with all the necessary components. Due to the presence of heterointerfaces and surfaces, there are spatially varying dielectric properties that can be described by a spatially varying relative dielectric constant $\epsilon(\mathbf{r})$. There will be fixed charges in the problem, such as ionized dopants or fixed surface charges, and electronic charges whose motion will have to be described within the effective mass Schrödinger equation (5.9) on page 74. Furthermore, there may be gate electrodes on surfaces on which voltages can be applied. The general treatment of this situation given in this chapter will enable us to gain further insight into the possibilities for creating confinement potentials for electrons (or holes), i.e., insight into the potential $U(\mathbf{r}_i)$ and the electron–electron interaction $V_C(\mathbf{r}_i - \mathbf{r}_j)$ in eq. (5.9).

The problem consists of finding the solution of Poisson's equation

$$\nabla\left[\epsilon(\mathbf{r})\epsilon_0\nabla\phi(\mathbf{r})\right] = -\rho(\mathbf{r}) \tag{7.1}$$

in which the electrostatic potential fulfills the following boundary conditions: on the surfaces S_i of the metallic electrodes (gates) the electro-

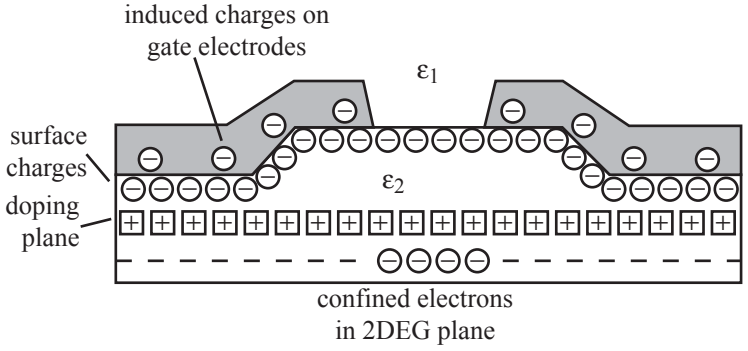


Fig. 7.1 Schematic illustration of a typical distribution of fixed and mobile charges in a semiconductor nanostructure with gate electrodes.

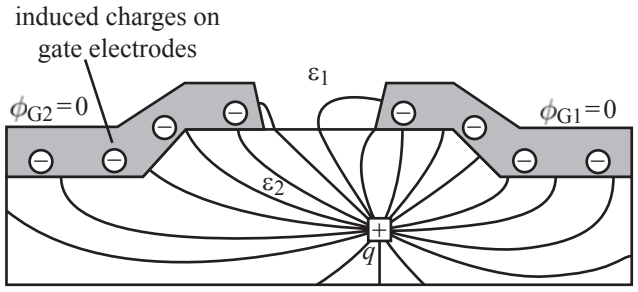


Fig. 7.2 Example of the potential described by Green's function, plotted as electric field lines. A unity charge $q = 1$ at $\mathbf{r}_1$ fills space with an electric field that is, at the gate electrodes, oriented normal to the surface. Surface charges will be induced on these surfaces. Electric field lines are deflected at dielectric interfaces.

static potential takes constant values ϕ_i, i.e.,

$$\phi(\mathbf{r})|_{S_i} = \phi_i.$$

The charge density $\rho(\mathbf{r})$ in eq. (7.1) describes the spatially fixed charges of ionized dopants, surface charges, or other ionized impurities, described by the density $\rho_{\text{ion}}(\mathbf{r})$, and the density of the mobile charge carriers (electrons) denoted as $\rho_{\text{e}}(\mathbf{r})$, such that $\rho(\mathbf{r}) = \rho_{\text{ion}}(\mathbf{r}) + \rho_{\text{e}}(\mathbf{r})$.

7.2 Formal solution using Green's function

An analytic solution of eq. (7.1) can only be found in very few special cases involving particular symmetries. Usually one has to find solutions numerically. However, there is a way of solving the problem formally by introducing Green's function. The solution of the problem will therefore express the electrostatic potential $\phi(\mathbf{r})$ with the help of this function. The different terms of this result can be physically interpreted, and insight into the problem can be gained even without actually calculating Green's function.

Green's function. We define Green's function $G(\mathbf{r}, \mathbf{r}_1)$ as the solution of the equation

$$\nabla \left[\epsilon(\mathbf{r})\epsilon_0 \nabla G(\mathbf{r}, \mathbf{r}_1)\right] = -\delta(\mathbf{r} - \mathbf{r}_1) \tag{7.2}$$

with the boundary conditions

$$G(\mathbf{r}, \mathbf{r}_1)|_{\mathbf{r}\in S_i} = 0.$$

Green's function describes the electrostatic potential at $\mathbf{r}$ *that is created by a unity point charge placed at* $\mathbf{r}_1$*, if all metal electrodes are grounded.* Green's function has the property $G(\mathbf{r}_1, \mathbf{r}_2) = G(\mathbf{r}_2, \mathbf{r}_1)$ (The proof is found in Appendix B).

Green's integral theorem. In order to solve eq. (7.1) formally, we further need an extended version of Green's integral theorem (the derivation

can be found in Appendix B):

$$\oint_S ds\,[\psi\epsilon\nabla\phi - \phi\epsilon\nabla\psi]\cdot\mathbf{n} = \int_V dV\,\{\psi\nabla\,[\epsilon\nabla\phi] - \phi\nabla\,[\epsilon\nabla\psi]\} \qquad (7.3)$$

Here the unit vector $\mathbf{n}$ is normal to surface S pointing outwards with respect to the volume enclosed by S.

General solution of the electrostatic problem. In order to solve our electrostatic problem, we replace in Green's theorem, eq. (7.3), ψ by Green's function $G(\mathbf{r},\mathbf{r}_1)$, ϵ by $\epsilon(\mathbf{r})\epsilon_0$, and ϕ by $\phi(\mathbf{r})$ from eq. (7.1), and obtain

$$\begin{aligned}\phi(\mathbf{r}) \;=\; & \int_V dV_1 G(\mathbf{r}_1,\mathbf{r})\rho_{\rm ion}(\mathbf{r}_1) \\ & + \int_V dV_1 G(\mathbf{r}_1,\mathbf{r})\rho_{\rm e}(\mathbf{r}_1) \\ & - \sum_i \phi_i \int\int_{S_i} ds_{1i}\,[\epsilon(\mathbf{r}_1)\epsilon_0\nabla_1 G(\mathbf{r}_1,\mathbf{r})]\cdot\mathbf{n}_i \qquad (7.4)\end{aligned}$$

This solves the electrostatic problem for a given charge density distribution, if Green's function $G(\mathbf{r}_1,\mathbf{r}_2)$ is known. *This function is given by the geometry of the gate electrodes and the relative dielectric function alone. In particular, it is independent of the voltages ϕ_i on the electrodes and of the charge distribution in the system.*

The solution for the potential in eq. (7.4) expresses the superposition principle. The potential appears to be the sum of various terms that can be interpreted individually. The first term describes the electrostatic potential created by the distribution of fixed charges for grounded gate electrodes and we define

$$\Phi_{\rm ion}(\mathbf{r}) := \int_V dV_1 G(\mathbf{r}_1,\mathbf{r})\rho_{\rm ion}(\mathbf{r}_1). \qquad (7.5)$$

The second term describes the electrostatic potential created by the distribution of electronic charges in the system and we define

$$\Phi_{\rm e}(\mathbf{r}) := \int_V dV_1 G(\mathbf{r}_1,\mathbf{r})\rho_{\rm e}(\mathbf{r}_1). \qquad (7.6)$$

The third term contains no charge density, but includes the voltages ϕ_i on the gate electrodes. It is a sum of contributions of all the individual gate electrodes. The characteristic potential of gate electrode i defined as

$$\alpha_i(\mathbf{r}) := -\int\int_{S_i} ds_{1i}\,[\epsilon(\mathbf{r}_1)\epsilon_0\nabla_1 G(\mathbf{r}_1,\mathbf{r})]\cdot\mathbf{n}_i \qquad (7.7)$$

describes the electrostatic potential distribution in the structure if there were no charges, with a unity voltage applied to gate i and zero voltage on all other gates. Another property of the characteristic potential $\alpha_i(\mathbf{r})$ is found if one considers the case in which all ϕ_i in eq. (7.4) have the

same value V_0. In this case we know that $\phi(\mathbf{r}) = V_0$ must hold and therefore it follows that

$$\sum_i \alpha_i(\mathbf{r}) = 1. \tag{7.8}$$

The solution of the electrostatic problem, eq. (7.4) can now be expressed in the form

$$\phi(\mathbf{r}) = \Phi_{\text{ion}}(\mathbf{r}) + \Phi_{\text{e}}(\mathbf{r}) + \sum_i \phi_i \alpha_i(\mathbf{r}). \tag{7.9}$$

7.3 Induced charges on gate electrodes

We will find an alternative interpretation of the characteristic potentials $\alpha_i(\mathbf{r})$ by calculating the screening charges induced on the gate electrodes. The total charge induced at the surface of electrode i is given by

$$\begin{aligned} Q_i &= \int\int_{S_i} dS_i \epsilon(\mathbf{r}) \epsilon_0 \nabla\phi(\mathbf{r}) \cdot \mathbf{n}_i \\ &= \int\int_{S_i} dS_i \epsilon(\mathbf{r}) \epsilon_0 \nabla \left\{ \Phi_{\text{ion}}(\mathbf{r}) + \Phi_{\text{e}}(\mathbf{r}) + \sum_j \phi_j \alpha_j(\mathbf{r}) \right\} \cdot \mathbf{n}_i. \end{aligned}$$

Again, the charge appears to be a superposition of different contributions. We can therefore rewrite the expression in the form

$$Q_i = Q_i^{(0)} + \sum_j C_{ij} \phi_j, \tag{7.10}$$

where we have defined the charge on gate electrode i for zero gate voltages, i.e., the screening charge on the gates induced by the static and electronic charges in the system,

$$Q_i^{(0)} := \int\int_{S_i} dS_i \epsilon(\mathbf{r}) \epsilon_0 \nabla \left\{ \Phi_{\text{ion}}(\mathbf{r}) + \Phi_{\text{e}}(\mathbf{r}) \right\} \cdot \mathbf{n}_i, \tag{7.11}$$

and the elements of the capacitance matrix

$$C_{ij} := \int\int_{S_i} dS_i \epsilon(\mathbf{r}) \epsilon_0 \nabla\alpha_j(\mathbf{r}) \cdot \mathbf{n}_i \tag{7.12}$$

describing the charges induced on the gates due to the application of voltages on them.

Screening charge induced on the gates by charges in the system. We can simplify the expression for $Q_i^{(0)}$ in eq. (7.11) by using the definitions of $\Phi_{\text{ion}}(\mathbf{r})$ and $\Phi_{\text{e}}(\mathbf{r})$ in eqs (7.5) and (7.6), and the definition of the characteristic potentials $\alpha_i(\mathbf{r})$ in eq. (7.7).

$$\begin{aligned} Q_i^{(0)} &= \int_V dV \left[\rho_{\text{ion}}(\mathbf{r}) + \rho_{\text{e}}(\mathbf{r}) \right] \int\int_{S_i} ds_{1i} \epsilon(\mathbf{r}_1) \epsilon_0 \nabla_1 G(\mathbf{r}, \mathbf{r}_1) \cdot \mathbf{n}_i \\ &= -\int_V dV \rho_{\text{ion}}(\mathbf{r}) \alpha_i(\mathbf{r}) - \int_V dV \rho_{\text{e}}(\mathbf{r}) \alpha_i(\mathbf{r}) \end{aligned} \tag{7.13}$$

Another meaning of the characteristic function $\alpha_i(\mathbf{r})$ emerges from this expression. If a charge $+e$ is placed at $\mathbf{r}$, the quantity $-e\alpha_i(\mathbf{r})$ is the screening charge induced on the surface of electrode i. From eq. (7.8) it follows that the sum over the charges induced on all gate electrodes is equal to $-e$.

Combining the two meanings of the characteristic function $\alpha_i(\mathbf{r})$—the electrostatic potential at $\mathbf{r}$, created by a unity voltage on electrode i, and the fraction of induced charge on electrode i, if a charge is placed at $\mathbf{r}$—allows us to find an important interpretation of the quantity $e\alpha_i(\mathbf{r})\phi_i$: On the one hand, this is the electrostatic energy of the charge e placed at $\mathbf{r}$, but on the other hand it is the work performed by the voltage source for bringing the screening charge $-e\alpha_i(\mathbf{r})$ onto electrode i.

The capacitance coefficients. The expression for the capacitance coefficients in eq. (7.12) can be rewritten by using eq. (7.7) and the result is

$$C_{ij} = \int\int_{S_i} ds_i \epsilon(\mathbf{r}_i)\epsilon_0 \int\int_{S_j} ds_j \epsilon(\mathbf{r}_j)\epsilon_0 \nabla_i \left\{[\nabla_j G(\mathbf{r}_i, \mathbf{r}_j)] \cdot \mathbf{n}_j\right\} \cdot \mathbf{n}_i.$$

These coefficients form the capacitance matrix of the system and they depend on the geometry of the system only. They are independent of the potentials on the gate electrodes.

The capacitance coefficients obey the equation

$$\sum_j C_{ij} = 0$$

because the charge on electrode i must not change, if all potentials ϕ_i are lifted by the same amount. Furthermore, they are symmetric in the indices i and j, i.e., $C_{ij} = C_{ji}$. This means that the number of independent matrix elements in a problem with N electrodes reduces to $N(N-1)/2$. Using these properties for the capacitance matrix of the system we obtain for the total induced charge on electrode i

$$Q_i = Q_i^{(0)} + \sum_j C_{ij} \left(\phi_j - \phi_i\right),$$

i.e., the induced charge depends only on voltage differences.

7.4 Total electrostatic energy

We are now interested in calculating the total electrostatic energy of our system. It is given by

$$W = \frac{1}{2}\int_V dV \left[\rho_{\text{ion}}(\mathbf{r}) + \rho_{\text{e}}(\mathbf{r})\right]\phi(\mathbf{r}) + \frac{1}{2}\sum_i \phi_i Q_i$$

The first term describes the energy of the charges within the system, and the second term is the energy of the screening charges induced on

the electrodes. Inserting the result for the potential $\phi(\mathbf{r})$, eq. (7.9), and for the screening charge Q_i induced on gate i, eq. (7.10), we obtain

$$W = W_{\text{ion-ion}} + W_{\text{ion-e}} + W_{\text{e-e}} + \frac{1}{2}\sum_{ij}\phi_i C_{ij}\phi_j, \tag{7.14}$$

where the energy of the fixed charges in the system has been defined to be

$$W_{\text{ion-ion}} := \frac{1}{2}\int dV \rho_{\text{ion}}(\mathbf{r})\Phi_{\text{ion}}(\mathbf{r}), \tag{7.15}$$

the energy of the electronic system is

$$W_{\text{e-e}} := \frac{1}{2}\int dV \rho_{\text{e}}(\mathbf{r})\Phi_{\text{e}}(\mathbf{r}), \tag{7.16}$$

and the interaction energy between the fixed charges and the electronic system is

$$W_{\text{ion-e}} := \int dV \rho_{\text{e}}(\mathbf{r})\Phi_{\text{ion}}(\mathbf{r}). \tag{7.17}$$

7.5 Simple model of a split-gate structure

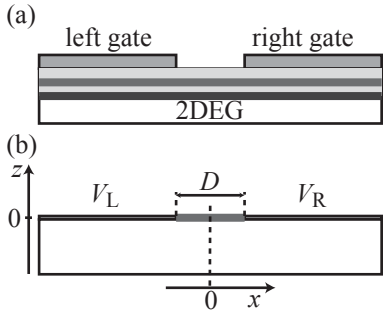


Fig. 7.3 Electrostatic model of a split gate on a heterostructure with remote doping and two-dimensional electron gas.

In order to illustrate the general considerations of the previous section, we discuss a simple model for a split-gate structure fabricated on top of a heterostructure in which a two-dimensional electron gas resides at the heterointerface below the surface (Glazman and Larkin, 1991). Figure 7.3(a) shows a cross-sectional view of such a structure. We are interested in the potential $\phi(x)$ in which the electrons move. This potential is, on the one hand, given by the fixed ionized dopants in the doping plane, and, on the other hand, by the voltages on the two gate electrodes. In accordance with the superposition principle expressed in eq. (7.9) we split the potential into two contributions. The first contribution, Φ_{G}, is caused by the voltages V_{L} and V_{R} on the two gate electrodes in the absence of the donor charges. The second contribution, Φ_{ion}, is created by the ionized donors, if $V_{\text{L}} = V_{\text{R}} = 0$. We assume that the split gate is much more extended in the y-direction than the width D of the slit between the two gates. In order to simplify the model, we neglect the separation of the plane of the gates, the donors and the electron gas as depicted in Fig. 7.3(b). This is a good approximation if these separations are small compared to the depletion lengths near the gate electrodes. In this model the semiconductor fills all space for which $z \leq 0$. We further assume that the donor charges at $|x| > D/2$ are completely neutralized by the image charges on the gate electrodes. As a result, only positive charges in the region $|x| < D/2$ are relevant.

The potential $\Phi_{\text{G}}(x, z)$ obeys Laplace's equation

$$\Delta\Phi_{\text{G}} = 0$$

with the boundary conditions

$$\Phi_{\text{G}}(x, z = 0) = \begin{cases} V_{\text{L}} \text{ for } x < -D/2 \\ V_{\text{R}} \text{ for } x > D/2 \end{cases} .$$

In agreement with eq. (7.9) the solution has the form

$$\Phi_{\mathrm{G}}(x,z) = V_{\mathrm{L}}\alpha_{\mathrm{L}}(x,z) + V_{\mathrm{R}}\alpha_{\mathrm{R}}(x,z)$$

and it can be shown (Glazman and Larkin, 1991) that the characteristic function $\alpha_{\mathrm{L}}(x,z)$ can be written as

$$\alpha_{\mathrm{L}}(x,z) \quad = \quad -\mathrm{sgn}(x)\frac{1}{\pi}\mathrm{Im}\ \ln\left\{\frac{2\zeta}{D} - \left[\left(\frac{2\zeta}{D}\right)^2 - 1\right]^{1/2}\right\},$$

where $\zeta = x + iz$, and $\alpha_{\mathrm{R}}(x,z) = 1 - \alpha_{\mathrm{L}}(x,z)$. The characteristic function $\alpha_{\mathrm{L}}(x,z)$ is plotted in Fig. 7.4.

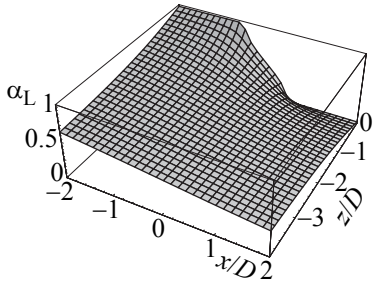


Fig. 7.4 Characteristic potential of the left gate, $\alpha_{\mathrm{L}}(x,z)$.

The potential $\Phi_{\mathrm{ion}}(x,z)$ due to the charged donors in the slit between the gates can be expressed as (Glazman and Larkin, 1991)

$$\Phi_{\mathrm{ion}}(x,z) = \frac{en_{\mathrm{s}}}{\varepsilon\varepsilon_0}\frac{D}{2}\mathrm{Im}\left\{\frac{2\zeta}{D} + i\left[1 - \left(\frac{2\zeta}{D}\right)^2\right]^{1/2}\right\},$$

where n_{s} is the electron density in the original two-dimensional electron gas. This potential normalized to $-en_{\mathrm{s}}D/2\varepsilon\varepsilon_0$ is plotted in Fig. 7.5. It fulfills the boundary condition $\Phi_{\mathrm{ion}}(x, z=0) = 0$ for $|x| > D/2$, i.e., on the metallic electrodes. This boundary condition implies that screening of the charged donors by the gate electrodes is accounted for, which has a significant influence on the potential. It is evident that the ions create an attractive potential for electrons with its minimum at $x = 0$. This potential is responsible for the confinement of the electron motion to a narrow channel. The steep potential walls of the confinement are a result of the screening effect of the gate electrodes.

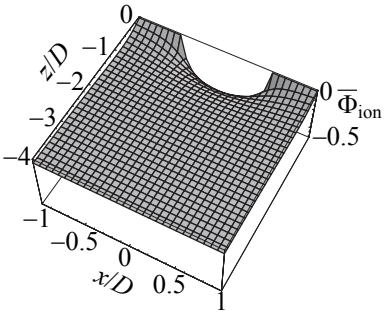


Fig. 7.5 Normalized potential of the donor layer $\bar{\Phi}_{\mathrm{ion}}(x,z)$.

The total potential of the system is the sum of $\Phi_{\mathrm{G}}(x,z)$ and $\Phi_{\mathrm{ion}}(x,z)$. If both gate voltages are the same, the gates contribute a constant to the total potential. If an asymmetric gate voltage is applied, i.e., $V_{\mathrm{L}} \neq V_{\mathrm{R}}$, the position of the potential minimum in the channel can be shifted, i.e., the electron channel is shifted in real space. In a real structure this can improve the situation if no ideal channel forms as a result of residual potential fluctuations due to the discreteness of the donor charges. In this case, asymmetrically applied gate voltages can frequently improve the situation.

In our considerations we have neglected the potential contribution that is caused by the electrons in the channel. This approximation is only valid if the channel is almost completely depleted. At high electron concentrations in the channel, this potential has to be taken into account self-consistently.

The question of how surface charges have to be included in the calculation is discussed in the literature. In experiments on nanostructures in GaAs/AlGaAs at temperatures below 4.2 K, it is frequently assumed that the surface charge distribution is frozen and does not change on the application of gate voltages. This is in contrast to models considering a pinned Fermi level at the surface, which implies a constant surface potential. In this case, charge will be exchanged between the electron gas

and the surface upon a change of gate voltages. This may be relevant at elevated temperatures. A comparison of these two models can be found in Davies *et al.* 1995.

Further reading

- Jackson 1983; Maxwell 1873.
- Papers: Glazman and Larkin 1991; Davies *et al.* 1995.

Exercises

(7.1) Consider a parallel plate capacitor with plate separation d and plate dimensions much larger than d. Both plates are grounded. Far away from the plate edges you place a single electron at a distance $x < d$ from one of the two plates. What is the charge induced on each of the two capacitor plates?

(7.2) Consider a parallel plate capacitor with plate separation d and plate dimensions much larger than d. One plate is kept grounded, the other is connected to a voltage source keeping it at the voltage $V_0 = 1\,\mathrm{V}$. A third, initially charge-neutral sheet of metal is inserted into this capacitor parallel to its two plates at a distance x ($0 < x < d$) to the grounded plate. What is the energy required to add a single electron (initially at ground potential) to the central plate? What is the energy to add a second electron?

(7.3) Two electrons are placed at the same distance d above a metallic plane connected to ground. Discuss how their mutual repulsion is altered by the presence of the metallic plane. How does the component parallel to the metallic plane of the repelling force between the electrons change with the separation x of the two electrons?

Quantum mechanics in semiconductor nanostructures

8

8.1 General hamiltonian

The electrons in a semiconductor move according to the predictions of Schrödinger's equation. The effective mass approximation leads to a Schrödinger equation for the envelope function. The many-particle problem of interacting particles is governed by a hamiltonian of the form shown in eq. (5.9). We will now specify the confinement and interaction potential in this hamiltonian taking screening effects of gate electrodes into account. We achieve this by starting from the total electrostatic energy of the system given in eq. (7.14) together with the definitions in eqs (7.15), (7.16), and (7.17).

Discrete electronic distribution. As a start, we specify the electron density distribution in our nanostructure as

$$\rho_{\mathrm{e}}(\mathbf{r}) = -|e| \sum_i \delta(\mathbf{r} - \mathbf{r}_i),$$

where the sum is over all electrons in the system and $\mathbf{r}_i$ describes the location of the ith electron. With this distribution, and eq. (7.5), we find for the electron–ion interaction energy, eq. (7.17), the expression

$$W_{\mathrm{ion-e}} = -|e| \sum_i \int_V dV' \rho_{\mathrm{ion}}(\mathbf{r}') G(\mathbf{r}', \mathbf{r}_i).$$

In a similar fashion we obtain for the electron–electron interaction energy, eq. (7.16), using eq. (7.6),

$$W_{\mathrm{e-e}} := \frac{e^2}{2} \sum_i G(\mathbf{r}_i, \mathbf{r}_i) + \frac{e^2}{2} \sum_{ij, j \neq i} G(\mathbf{r}_i, \mathbf{r}_j).$$

The second term describes the mutual interaction between electrons. The first term is a self-interaction term. For a system without dielectric interfaces and gate electrodes, this term would diverge and should be neglected. However, if gate electrodes and dielectric interfaces are present, this term also describes the interaction between an electron and its image charges. We will therefore keep this term in the following.

Potential energy of an N-electron system. While eq. (7.14) is an expression for the total electrostatic energy of a system, we are now interested in the electrostatic energy of the electronic system alone. In order to find this energy, we determine how much energy we have to use to build up the electronic system. Assume that there are already $p-1$ electrons in the system and we wish to add the pth electron. The required energy is

$$\begin{aligned}\Delta W_p &= W_p - W_{p-1} \\ &= -|e|\int_V dV \rho_{\rm ion}(\mathbf{r})G(\mathbf{r}_p,\mathbf{r}) + \frac{e^2}{2}G(\mathbf{r}_p,\mathbf{r}_p) + e^2\sum_{n=1}^{p-1} G(\mathbf{r}_n,\mathbf{r}_p)\end{aligned}$$

A part of this energy is, however, provided by the voltage sources connected to the gate electrodes, because when we add the electron the induced screening charges on the gates change as well. This change in charge is, according to eqs (7.10) and (7.13) given by

$$\Delta Q_{i,p} = |e|\alpha_i(\mathbf{r}_p).$$

The work done by the voltage sources is then

$$\Delta W_{p,\rm sources} = \sum_i \phi_i \Delta Q_{i,p} = |e|\sum_i \phi_i\alpha_i(\mathbf{r}_p).$$

The required extra energy to add the pth electron is the difference $\Delta W_p - \Delta W_{p,\rm sources}$, i.e.,

$$\begin{aligned}V_p(\mathbf{r}_p) &= -|e|\int_V dV\rho_{ion}(\mathbf{r})G(\mathbf{r}_p,\mathbf{r}) \\ &\quad + \frac{e^2}{2}G(\mathbf{r}_p,\mathbf{r}_p) + e^2\sum_{n=1}^{p-1}G(\mathbf{r}_n,\mathbf{r}_p) \\ &\quad - |e|\sum_i \phi_i\alpha_i(\mathbf{r}_p).\end{aligned}$$

This expression can be interpreted as the potential energy of the pth electron. We obtain the total potential energy of the electronic system by summing $V_p(\mathbf{r}_p)$ over all N electrons of the system.

Hamiltonian for the N-electron system. We are now ready to write down the hamiltonian for the N-electron system in the effective mass approximation [cf., eq. (5.9)] as

$$\begin{aligned}H_N &= \sum_{n=1}^{N}\left\{\frac{[\mathbf{p}_n + |e|\mathbf{A}(\mathbf{r}_n)]^2}{2m^\star} + U(\mathbf{r}_n) + \frac{1}{2}g^\star\mu_{\rm B}\boldsymbol{\sigma}_n\mathbf{B}(\mathbf{r}_n)\right. \\ &\quad \left. + e^2\sum_{m=1}^{n-1}G(\mathbf{r}_m,\mathbf{r}_n)\right\},\end{aligned} \tag{8.1}$$

where

$$U(\mathbf{r}) = E_c(\mathbf{r}) - |e| \int_V dV' \rho_{\text{ion}}(\mathbf{r}') G(\mathbf{r}, \mathbf{r}') + \frac{e^2}{2} G(\mathbf{r}, \mathbf{r}) - |e| \sum_i \phi_i \alpha_i(\mathbf{r}) \tag{8.2}$$

This hamiltonian is valid in materials with a parabolic and isotropic band in which the wave function is concentrated mainly in one material. The first term in eq. (8.1) describes the kinetic energy of the electrons and allows for a magnetic field. The second term describes a single-particle confinement potential originating from the terms specified in eq. (8.2). The first stems from a change in material composition. Such a change can be abrupt, as in a quantum well, or continuous, as in a parabolic quantum well. The second term in eq. (8.2) is the potential caused by the distribution of fixed charges in the system, such as ionized dopants, ionized impurities, and fixed surface charges. Often the discreteness of the charge distribution is neglected and an average charge density per volume is used instead (Jellium model). The third term in eq. (8.2) describes the interaction of an electron with its own screening charges induced on the gate electrodes and at dielectric interfaces. Its diverging self-energy contribution has to be removed by suitable mathematical techniques. The action of voltages ϕ_i applied to the gates is given by the last term in eq. (8.2). In the hamiltonian (8.1) the third term, also called the Zeeman term, acts on the spin-degree of freedom and leads to an energy splitting of spin-degenerate levels if the magnetic field is finite. The last term in (8.1) describes the electron–electron interaction. It is responsible for the many-body nature of the problem. It is frequently considered within approximations, such as the Hartree and the Hartree-Fock methods, or density-functional theory.

The above hamiltonian (8.1) nicely summarizes the ways in which we can tailor the confinement potential for electrons in semiconductor nanostructures. The choice of the primary material determines the effective mass $m^\star$ and the effective g-factor $g^\star$. Combining different materials in heterostructures leads to band offsets acting as an effective confinement potential in (8.2), but also to modifications of electronic interactions, if the heterointerface is a dielectric interface [boundary conditions for $G(\mathbf{r}, \mathbf{r}')$]. The distribution of doping atoms in the structure contributes to the confinement, but also causes spatial fluctuations of the potential at the location of the electronic states due to the discreteness of the distribution. The same is valid for residual charged background impurities in the material. Significant freedom is given for tailoring the confinement potential via the last term in (8.2) with suitable gate geometries and by applying appropriate gate voltages. This term is responsible for the confinement of electrons in laterally patterned split-gate devices, such as quantum point contacts and quantum dots. At the same time, gate electrodes tend to screen the interaction among electrons and to flatten spatial potential fluctuations of charged impurities by imposing boundary conditions on Green's function $G(\mathbf{r}, \mathbf{r}')$.

Example: screening in a two-dimensional electron gas with gate. In order to demonstrate the screening effect due to the presence of a gate electrode, we consider a two-dimensional electron gas in a GaAs/AlGaAs heterostructure with top gate. The electron gas is at a distance d parallel to the gate. We assume for simplicity that GaAs and $Al_xGa_{1-x}As$ have exactly the same relative dielectric constant ε. The z-axis is the growth direction and therefore all interfaces and the plane of the electron gas are at $z = \text{const}$. Green's function for this system is given by

$$G(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{4\pi\varepsilon\varepsilon_0}\left(\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} - \frac{1}{|\mathbf{r}_1 - \mathbf{r}_3|}\right),$$

where $\mathbf{r}_3 = (\mathbf{r}_2\mathbf{e}_x)\mathbf{e}_x + (\mathbf{r}_2\mathbf{e}_y)\mathbf{e}_y - (\mathbf{r}_2\mathbf{e}_z)\mathbf{e}_z$. We find for the interaction between electrons in the plane of the electron gas the expression

$$G(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{4\pi\varepsilon\varepsilon_0\rho}\left(1 - \frac{1}{\sqrt{1 + \left(\frac{2d}{\rho}\right)^2}}\right),$$

where ρ is the separation of the two electrons in the plane. For small separations $\rho \ll 2d$ the interaction between electrons is essentially not modified by the presence of the gate and is proportional to $1/\rho$. For $\rho \gg 2d$ the interaction decays proportional to $1/\rho^3$. Each electron experiences the field of an electric dipole created by the other electron and its image charge. This is how the gate screens the electron–electron interaction.

8.2 Single-particle approximations for the many-particle problem

The Schrödinger problem with the hamiltonian (8.1) cannot be solved analytically. Numerical algorithms have to be applied and clever approximation methods have to be used. There are a number of such approximations that are quite commonly applied to many-body problems. Among them are the local density approximation, the Hartree approximation, the Hartree–Fock approximation, and the Thomas–Fermi approximation.

Local density approximation. Density-functional theory (DFT) was introduced by Hohenberg, Kohn and Sham (Hohenberg and Kohn, 1964; Kohn and Sham, 1965). In 1998, Walter Kohn was awarded the Nobel prize in chemistry for his development of the density-functional theory. It expresses the total energy of an interacting system as a functional of the electron density. The theory was developed for the case where Green's function is given by the bare Coulomb potential, i.e., the case in which image charge effects due to gates and dielectric interfaces are not relevant. In this case, the third term in eq. (8.2) drops out and the

total energy of the system can (in units of $E^\star_{\rm Ry}$ and $a^\star_{\rm B}$) be written as (Hohenberg and Kohn, 1964; Kohn and Sham, 1965)

$$\begin{aligned} E_v[n] &= \frac{1}{2}\int dV \nabla_{\mathbf{r}}\nabla_{\mathbf{r}'}\, n_1(\mathbf{r},\mathbf{r}')|_{\mathbf{r}=\mathbf{r}'} + \int dV U(\mathbf{r}) n(\mathbf{r}) \\ &\quad + \frac{1}{2}\int\int dV dV' \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + E_{\rm xc}[n(\mathbf{r})]. \end{aligned} \tag{8.3}$$

Here, $n_1(\mathbf{r},\mathbf{r}')$ is the single-particle density matrix. The first term describes the kinetic energy of the system corresponding to the first term of (8.1) with zero magnetic field, the second is the potential energy in the external potential $U(\mathbf{r})$, given by eq. (8.2). The third and fourth terms are contributions to the interaction, i.e., the last term in the hamiltonian (8.1), the third describing the classical Coulomb energy, and $E_{\rm xc}[n(\mathbf{r})]$ being the so-called exchange–correlation energy. The interaction terms can be written down in a straightforward way for the case of gated structures with Green's function replacing the bare Coulomb interaction. It has been shown, that the correct ground state electron density distribution minimizes the energy functional $E_v[n(\mathbf{r})]$, if $n(\mathbf{r})$ fulfills the condition

$$\int dV n(\mathbf{r}) = N,$$

where N is the number of electrons in the system. So far this formalism does not employ any approximations.

The problem, however, is to find an explicit expression of the exchange–correlation energy functional. In the *local density approximation (LDA)*, which is valid in inhomogeneous electron gases with slowly varying density, one can approximate

$$E_{\rm xc}[n(\mathbf{r})] = \int dV n(\mathbf{r}) \epsilon_{\rm xc}[n(\mathbf{r})],$$

where $\epsilon_{\rm xc}[n]$ is the exchange–correlation energy of a single electron in a homogeneous electron gas of constant density n.

Minimization of the energy functional $E_v[n]$ in eq. (8.3) leads in this approximation to a self-consistent single-electron Schrödinger equation of the form (Kohn and Sham, 1965)

$$\left\{ \frac{\mathbf{p}^2}{2m^\star} + U(\mathbf{r}) + V_{\rm H}(\mathbf{r}) + V_{\rm xc}[n(\mathbf{r})] \right\} \psi_n(\mathbf{r}) = E_n \psi_n(\mathbf{r}), \tag{8.4}$$

where $V_{\rm xc}[n] = d[n\epsilon_{\rm xc}(n)]/dn$. For the potential $U(\mathbf{r})$ we can explicitly write

$$U(\mathbf{r}) = E_{\rm c}(\mathbf{r}) - |e| \int_V dV' \rho_{\rm ion}(\mathbf{r}') G(\mathbf{r},\mathbf{r}') - |e| \sum_i \phi_i \alpha_i(\mathbf{r}). \tag{8.5}$$

The Hartree potential $V_{\rm H}(\mathbf{r})$ is a solution of Poisson's equation

$$\Delta V_{\rm H} = \frac{e^2 n(\mathbf{r})}{\varepsilon\varepsilon_0}, \tag{8.6}$$

with appropriate boundary conditions, and the density $n(\mathbf{r})$ is determined from the wave functions via

$$n(\mathbf{r}) = \sum_{n=1}^{N} |\psi_n(\mathbf{r})|^2 ,$$

where the sum extends over the N energetically lowest states. Effects of finite temperature can be taken into account by taking the distribution of electrons according to the Fermi–Dirac distribution into account. In this case the density is given by

$$n(\mathbf{r}) = \sum_{n=1}^{N} |\psi_n(\mathbf{r})|^2 \, f(E_n - E_\mathrm{F}), \tag{8.7}$$

and the Fermi energy has to be determined from the requirement of local charge neutrality

$$\int dV \, [\rho_\mathrm{ion}(\mathbf{r}) - |e| n(\mathbf{r})] = 0. \tag{8.8}$$

Approximate expressions for $V_\mathrm{xc}[n]$ assuming an unscreened Coulomb interaction potential have been given by a number of authors (Gunnarsson and Lundqvist, 1976; Hedin and Lundqvist, 1971; Perdew and Zunger, 1981; Roesler *et al.*, 1984). For example, the form of Gunnarsson and Lundqvist is (in units of $E^\star_\mathrm{Ry}$)

$$V_\mathrm{xc}[r_\mathrm{s}] = -\left[1 + 0.0545 r_\mathrm{s} \ln\left(1 + \frac{11.4}{r_\mathrm{s}}\right)\right] \frac{2}{\pi \alpha r_\mathrm{s}}, \tag{8.9}$$

with the interaction parameter

$$r_\mathrm{s} = \left[\frac{4\pi}{3} {a^\star_\mathrm{B}}^3 n\right]^{-1/3} .$$

Equations (8.4–8.9) are the set of equations that have to be solved self-consistently. The total minimized energy of the whole system is then given by

$$E = \sum_n \left(\langle T_n \rangle + \langle U_n \rangle + \frac{1}{2} \langle V_\mathrm{H}^{(n)} \rangle + \langle V_\mathrm{xc}^{(n)} \rangle \right), \tag{8.10}$$

where the expectation values $\langle \ldots \rangle$ are summed over all occupied single-particle states, and n denotes the quantum numbers. The quantities $\langle T_n \rangle$, $\langle U_n \rangle$, $\langle V_\mathrm{H}^{(n)} \rangle$, and $\langle V_\mathrm{xc}^{(n)} \rangle$ denote expectation values of the kinetic, the potential, the Hartree and the exchange–correlation energies, respectively. We emphasize here that the factor $1/2$ in front of $\langle V_\mathrm{H}^{(n)} \rangle$ stems from eq. (8.3) and avoids double counting of the electron–electron interaction terms. By contrast, the energy eigenvalues E_n obtained from the single-particle eq. (8.4) are given by

$$E_n = \langle T_n \rangle + \langle U_n \rangle + \langle V_\mathrm{H}^{(n)} \rangle + \langle V_\mathrm{xc}^{(n)} \rangle \tag{8.11}$$

without the factor $1/2$.

Hartree approximation. The Hartree approximation emerges from the local density approximation by neglecting the exchange–correlation potential $V_{\rm xc}(\mathbf{r})$ in eq. (8.4). The many-body wave function $\Psi(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ is approximated as the product of N single-particle wave functions $\psi(\mathbf{r}_i)$. The latter are chosen such that the total energy of the system is minimized. The resulting single-particle Schrödinger equation reads

$$\left\{ \frac{\mathbf{p}^2}{2m^\star} + U(\mathbf{r}) + V_{\rm H}(\mathbf{r}) \right\} \psi_n(\mathbf{r}) = E_n \psi_n(\mathbf{r}), \tag{8.12}$$

where $U(\mathbf{r})$ is determined by eq. (8.5), the Hartree potential is the solution of eq. (8.6) with the density given by eq. (8.7), and the Fermi energy determined by eq. (8.8).

Example I: Electron distribution in delta-doped layers. As a first example for the application of the self-consistent method of calculating the electronic structure of nanostructures we reconsider the case of the delta-doped layer introduced on page 73. We employ the Hartree approximation where each individual electron moves in the effective potential $U(\mathbf{r}) + V_{\rm H}(\mathbf{r})$, where the Hartree potential $V_{\rm H}(\mathbf{r})$ is the electrostatic potential created by all other electrons. In case of the delta-doped layer, no heterointerface and no gate electrodes are involved. Therefore we have in eq. (8.5) $E_{\rm c}(\mathbf{r}) = 0$ and $\alpha_i(\mathbf{r}) = 0$. The second term in eq. (8.5) is given by eq. (5.8), if we regard the doping plane as being uniformly charged thereby neglecting the spatial discreteness of the ionized donors. As a result, this problem reduces to the one-dimensional Schrödinger equation

$$\left[-\frac{\hbar^2}{2m^\star} \frac{\partial^2}{\partial z^2} + \frac{e^2 N_{\rm D}}{2\varepsilon\varepsilon_0} |z| + V_{\rm H}(z) \right] F_n(z) = E_n F_n(z). \tag{8.13}$$

The Hartree potential is found from the solution of Poisson's equation

$$\frac{d^2 V_{\rm H}(z)}{dz^2} = -\frac{e^2 n(z)}{\varepsilon\varepsilon_0}. \tag{8.14}$$

The electron density in this equation can be determined from the envelope functions via

$$\begin{aligned} n(z) &= 2 \cdot \sum_{n,\mathbf{k}} |F_n(z)|^2 \, \frac{1}{e^{(E_{n\mathbf{k}} - \mu)/kT} + 1} \\ &= 2kT \cdot \frac{m^\star}{\pi\hbar^2} \cdot \sum_n |F_n(z)|^2 \ln\left(e^{(\mu - E_n)/kT} + 1 \right). \end{aligned} \tag{8.15}$$

In the last step we have assumed a parabolic conduction band dispersion. The chemical potential μ in eq. (8.15) has to be chosen such that charge neutrality is established, meaning that

$$\int_{-\infty}^{+\infty} dz n(z) = N_{\rm D}. \tag{8.16}$$

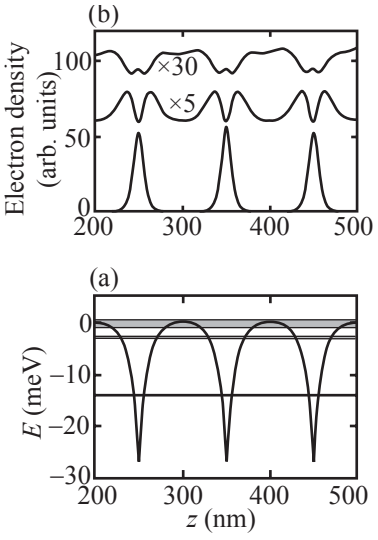


Fig. 8.1 (a) Self-consistently calculated conduction band edge and bound energy levels of three δ-doped layers in GaAs. (b) Electron densities of the three bound subband states. (Reprinted with permission from Kostial *et al.*, 1993. Copyright 1993 by the American Physical Society.)

Equations (8.13)–(8.16) are the system of equations for determining the bound states of the δ-doped layer that has to be solved self-consistently.

The result of such a calculation for three δ-doped layers in GaAs with a spacing of 100 nm is shown in Fig. 8.1. The electrons that are bound to the doping plane screen the potential of the δ-layers such that, from a distance, the system looks charge-neutral. This situation is similar to atoms, where the strong internal electric fields caused by the charged nucleus are screened by the electrons such that, from a distance, the atom is seen as a charge-neutral object.

The self-consistent Hartree method for solving Schrödinger's equation, Poisson's equation, and the charge density equation under the requirement of charge neutrality is very powerful and can also be applied to systems without translational invariance in the plane. It describes the interplay between quantization and screening that is very important in many semiconductor nanostructures.

Example II: Formation of a one-dimensional channel in a split-gate structure As the second example for the application of the Hartree approximation we discuss a calculation by Laux, Frank, and Stern (Laux *et al.*, 1988). They investigated the formation of a one-dimensional conducting channel in the split-gate structure shown schematically in Fig. 8.2. At zero gate voltages, a two-dimensional electron gas exists at the interface between GaAs and AlGaAs. If a negative voltage is applied to the gates, the electron gas under the gates is depleted and a narrow electronic channel develops below the slit between the two gates. The problem is translationally invariant in the y-direction. Therefore, in Schrödinger's equation, the y-direction separates.

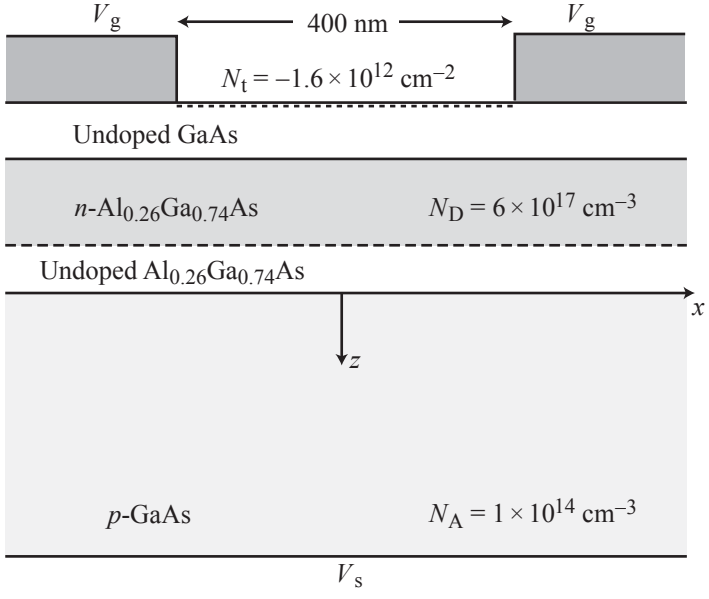


Fig. 8.2 Schematic cross-section of the structure for which the self-consistent calculation of the electronic channel is made. The structure is a GaAs/AlGaAs heterostructure with modulation doping and split gate. (Reprinted from Laux *et al.*, 1988 with permission from Elsevier.)

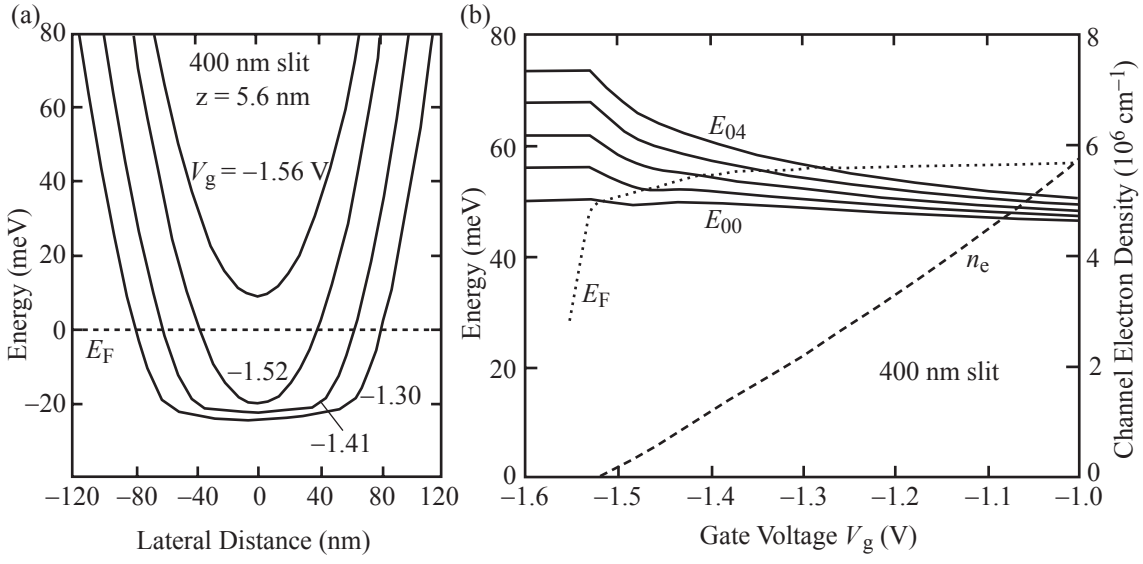


Fig. 8.3 (a) Effective confinement potential $U(\mathbf{r}) + V_H(\mathbf{r})$ in the x-direction in the plane of the two-dimensional electron gas. (b) Energy levels and Fermi level (electrochemical potential) for electrons. (Reprinted from Laux *et al.*, 1988 with permission from Elsevier.)

We write for the total wave function

$$\psi_{n\mathbf{k}}(\mathbf{r}) = \chi_n(x,z)e^{ik_y y}$$

and for the total energy

$$E_n(k_y) = E_n + \frac{\hbar^2 k_y^2}{2m^\star}.$$

The quantum number n is discrete, while k_y is a continuous variable.

Figure 8.3(a) shows the total potential $U(x,z) + V_H(x,z)$ along x in the plane of the electron gas calculated self-consistently for a number of gate voltages. The zero of energy is the Fermi energy. The calculation was made for a temperature $T < 4.2$ K. Bound quantum states arise in the x-direction in the almost parabolic potential. They are called one-dimensional modes that are described by the wave functions $\chi_n(x,z)$ and the corresponding energies E_n. At a gate voltage of -1.56 V the whole potential is above the Fermi energy and no states are occupied. The channel is depleted. At a gate voltage of -1.52 V the shape of the confinement potential is still parabolic, but its minimum is below the Fermi energy. In order to find out whether the lowest quantum state E_0 is occupied, we have to examine the quantization energies. These are plotted in Fig. 8.3(b). It turns out that above -1.52 V the lowest bound state is occupied. This is also evident in the plots of the electron density in Fig. 8.4. The electron density has the shape of the squared ground state wave function.

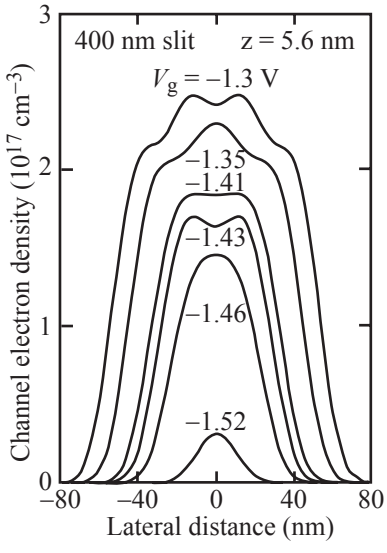


Fig. 8.4 Electron density at a number of gate voltages. The occupation of subbands shows in a modulation of the electron density normal to the channel direction. (Reprinted from Laux *et al.*, 1988 with permission from Elsevier.)

In Fig. 8.3(b) we can see that more and more modes become occupied with increasing gate voltage. As a result, the confinement potential gets broader and forms a flat bottom [see Fig. 8.3(a)]. Accordingly, the electron density in Fig. 8.4 becomes broader and shows an oscillatory structure caused by the contributions of higher subbands. The energetic separation of the states depends on the gate voltages and is between 1 and 5 meV corresponding to a temperature of about 11 K. For the observation of these quantization effects, experiments at low temperatures have to be performed.

Hartree–Fock approximation. In the Hartree–Fock approximation the many-body wave function is taken to be a Slater-determinant of single-particle wave functions. This accounts for the fermionic character of the electrons. Minimization of the total energy results in the Hartree equation (8.12) with an additional term describing the exchange interaction. While this term can be accounted for explicitly in nanostructures with a small number of electrons, e.g., in few-electron quantum dots, in large systems the local density approximation with its local exchange–correlation potential eq. (8.9) can be applied.

Thomas–Fermi approximation. Within the approximation of Thomas and Fermi the necessity of solving Schrödinger's equation can be completely avoided by also expressing the kinetic energy in eq. (8.4) as a local functional of the density. Essentially, one determines the local electron density by filling the three-dimensional density of states $\mathcal{D}(E)$ (in two-dimensional problems the 2D density of states) from the local band edge (subband edge) $E_0(\mathbf{r}) = U(\mathbf{r}) + V_\mathrm{H}(\mathbf{r})$ up to the Fermi energy:

$$\rho_e(\mathbf{r}) = -|e| \int dE \mathcal{D}[E - U(\mathbf{r}) - V_\mathrm{H}(\mathbf{r})] f(E - E_\mathrm{F}).$$

This electron density is plugged into Poisson's eq. (8.6) for determining V_H. This method is particularly useful if the potential $U(\mathbf{r})$ has small variations on the scale of the Fermi energy on length scales large compared to the Fermi-wavelengths of the electron. The quantization of states is completely neglected.

Further reading

- Density-functional theory: Giuliani and Vignale 2005.
- Hartree and Hartree–Fock approximations: Landau and Lifschitz 1962; Madelung 1972; Giuliani and Vignale 2005.
- Thomas–Fermi approximation: Kittel 2005; Ashcroft and Mermin 1987; Giuliani and Vignale 2005.
- Papers: Laux *et al.* 1988.

Exercises

(8.1) (a) Consider a shallow, low-density, two-dimensional electron gas in a Ga[Al]As heterostructure with top gate. The separation between the plane of the electron gas and the gate is $d = 34\,\mathrm{nm}$, the electron density is $n_\mathrm{s} = 1 \times 10^{11}\,\mathrm{cm}^{-2}$, and the relative dielectric constant of GaAs is $\varepsilon = 12.8$. Discuss the importance of screening by the top gate for the electron–electron interaction in the two-dimensional electron gas.

(b) Consider the same structure without a top gate. How does the electron–electron interaction potential change?

(8.2) Consider a single ionized donor with positive charge $+e$ located in the center of an $Al_{0.3}Ga_{0.7}As$ barrier of 34 nm thickness. On one side of the barrier there is a metallic gate electrode; on the other side there is GaAs. Write down an expression for the potential which an electron near the GaAs/AlGaAs heterointerface would see. How deep is the potential and what is its extent in the plane of the interface?

This page intentionally left blank

Two-dimensional electron gases in heterostructures

9

The physics of two-dimensional electron gases is very rich and interesting. Furthermore, two-dimensional electron gases in heterostructures are fundamental building blocks of semiconductor nanostructures. A large number of high quality semiconductor nanostructures have been made by lateral patterning. We will therefore apply the general techniques described in the previous chapters within a case study of GaAs/AlGaAs heterostructures.

9.1 Electrostatics of a GaAs/AlGaAs heterostructure

Consider a GaAs/AlGaAs heterostructure as it is depicted schematically in Fig. 9.1. As a first step we are interested in the electrostatic description of this structure. For simplicity we assume that the relative dielectric constants of GaAs and AlGaAs are identical. We choose the z-axis in the growth direction of the crystal, normal to the heterointerface with its origin, $z = 0$, at this interface. The AlGaAs barrier material is in the region $z < 0$, GaAs fills the half space $z > 0$. On top of the GaAs cap layer a thick metal layer has been deposited (but not shown in the figure).

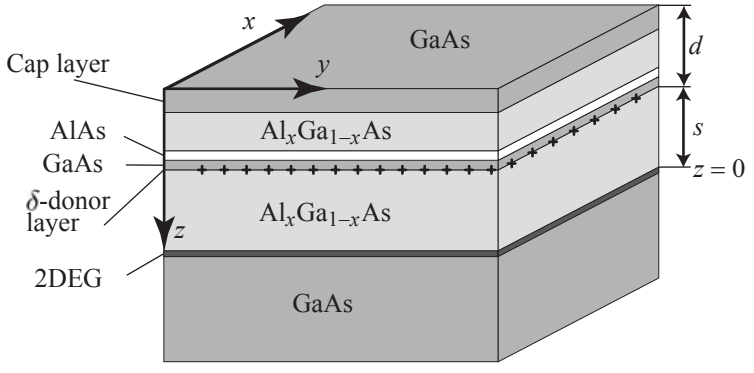


Fig. 9.1 Layer sequence in a typical GaAs/AlGaAs heterostructure with remote doping.

Jellium model. Doping atoms are randomly placed within the doping plane. We describe their distribution as

$$\mathcal{N}_\mathrm{d}(\mathbf{r},z)=\sum_i \delta(\mathbf{r}-\mathbf{r}_i)\delta(z+s),$$

where $\mathbf{r}=(x,y)$ and the $\mathbf{r}_i$ denote the positions of doping atoms in the plane. The average doping density is given by $\langle\mathcal{N}_\mathrm{d}(\mathbf{r},z)\rangle = N_\mathrm{d}\delta(z+s)$. The discrete distribution of doping atoms is often dealt with by splitting it into a constant $\mathbf{r}$-independent spatially averaged distribution and a fluctuating part:

$$\mathcal{N}_\mathrm{d}(\mathbf{r},z)=N_\mathrm{d}\delta(z+s)+C(\mathbf{r})\delta(z+s).$$

This equation defines the fluctuating part $C(\mathbf{r})$. Its spatial average vanishes, i.e., $\langle C(\mathbf{r})\rangle = 0$.

As a result of the superposition principle, the electrostatic potential created by the charged dopants can similarly be split into two contributions. The first is caused by the mean doping density $N_\mathrm{d}\delta(z+s)$. It is independent of $\mathbf{r}$, but depends on z. The second contribution results from the fluctuating part of the doping density $C(\mathbf{r})\delta(z+s)$. It leads to a spatially fluctuating potential with zero spatial average.

Within the jellium model, the spatially fluctuating part $C(\mathbf{r})$ of the distribution of dopants is neglected. As a result the problem becomes translationally invariant in the (x,y)-plane simplifying the electrostatics and quantum mechanics considerably. Building on the solution of the jellium model the fluctuations $C(\mathbf{r})$ can later be introduced within two-dimensional screening and scattering theory.

Electrostatics within the jellium model. For $z \gg 0$ the electric field in the sample is zero and the conduction band edge is flat. If we place a cylindrical closed surface along z with one end face in the region $z \gg 0$ and the other in the region $-s < z < 0$, we can apply Gauss's law of electrostatics and find the electric field in the spacer layer

$$E=\frac{|e|n_\mathrm{s}}{\varepsilon\varepsilon_0},$$

and the corresponding electrostatic potential

$$\phi(z)=-\frac{|e|n_\mathrm{s}}{\varepsilon\varepsilon_0}z \quad \text{for } -s<z<0.$$

If we extend the cylinder further in the negative z-direction, we include the δ-doping layer and find the new value

$$E=\frac{|e|(n_\mathrm{s}-N_\mathrm{d})}{\varepsilon\varepsilon_0},$$

and correspondingly

$$\phi(z)=\frac{|e|n_\mathrm{s}}{\varepsilon\varepsilon_0}s-\frac{|e|(n_\mathrm{s}-N_\mathrm{d})}{\varepsilon\varepsilon_0}(z+s) \quad \text{for } -s-d<z<-s.$$

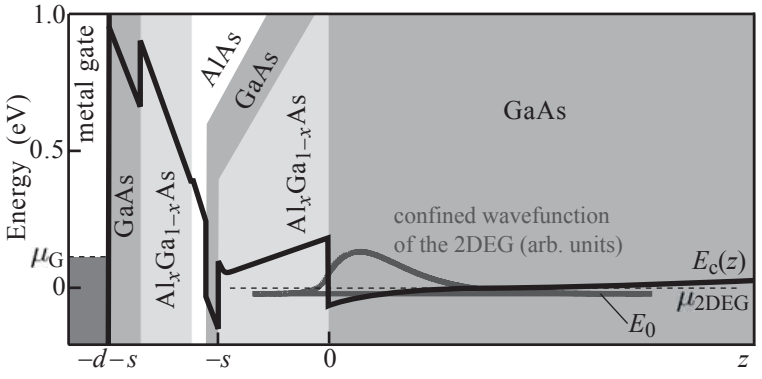


Fig. 9.2 Effective potential for electrons in the conduction band in a typical GaAs/AlGaAs heterostructure with remote doping.

At the semiconductor/metal interface the potential takes the value

$$\phi(-s-d) = \frac{|e|n_\mathrm{s}}{\varepsilon\varepsilon_0}s + \frac{|e|(n_\mathrm{s}-N_\mathrm{d})}{\varepsilon\varepsilon_0}d.$$

The effective potential for electrons in the conduction band is given by $E_\mathrm{c}(z) = -|e|\phi(z)$, where we choose $E_\mathrm{c}(z) = 0$ in GaAs and the conduction band offset ΔE_c in AlGaAs. The total potential is shown in Fig. 9.2.

9.2 Electrochemical potentials and applied gate voltage

Owing to Fermi level pinning at the metal/GaAs interface the electrochemical potential (Fermi level) at the surface is at the energy

$$\mu_\mathrm{G} = E_\mathrm{c}(-s-d) - \Phi_\mathrm{b},$$

where Φ_b is the built-in potential which is about half the band gap. Within the electron gas the electrochemical potential is given by the sum of the (not yet known) quantization energy and the Fermi energy:

$$\mu_\mathrm{2DEG} = E_0(n_\mathrm{s}) + E_\mathrm{F}(n_\mathrm{s}).$$

As a consequence, the relation between an applied gate voltage U_G between top gate and electron gas is

$$-|e|U_\mathrm{G} = \mu_\mathrm{G} - \mu_\mathrm{2DEG} = -\frac{e^2 n_\mathrm{s}}{\varepsilon\varepsilon_0}s - \frac{e^2(n_\mathrm{s}-N_\mathrm{d})}{\varepsilon\varepsilon_0}d - \Phi_\mathrm{b} - E_0(n_\mathrm{s}) - E_\mathrm{F}(n_\mathrm{s}).$$

9.3 Capacitance between top gate and electron gas

From this relation, the capacitance per unit area between top gate and electron gas can be calculated to be

$$\frac{1}{C/A} = -\frac{d(-|e|U_{\mathrm{G}})}{e^2 dn_{\mathrm{s}}} = \frac{1}{\varepsilon\varepsilon_0}\left(s + d + \frac{\varepsilon\varepsilon_0}{e^2}\frac{dE_0(n_{\mathrm{s}})}{dn_{\mathrm{s}}} + \frac{\varepsilon\varepsilon_0}{e^2}\frac{dE_{\mathrm{F}}(n_{\mathrm{s}})}{dn_{\mathrm{s}}}\right). \tag{9.1}$$

In addition to the simple geometrical contribution to the capacitance which is given by the separation $s + d$ between metal and heterointerface, there is a quantum capacitance contribution expressed by the last two terms in brackets. We identify the quantity $dE_{\mathrm{F}}(n_s)/dn_{\mathrm{s}}$ with the inverse of the system's density of states $\mathcal{D}_{2\mathrm{D}} = m^\star/\pi\hbar^2$. The corresponding length scale is $\varepsilon\varepsilon_0/e^2\mathcal{D}_{2\mathrm{D}} = a_{\mathrm{B}}^\star/4$. The finite density of states in the electron gas increases the effective separation of the capacitor plates by $a_{\mathrm{B}}^\star/4$, which gives, in GaAs, about 2.5 nm. The quantity $C_{\mathrm{q}} := e^2\mathcal{D}_{2\mathrm{D}}A$ has the dimensions of a capacitance and is called quantum capacitance. The fact that the density of states of a two-dimensional system enters the capacitance is exploited in the capacitance spectroscopy method. For example, in a magnetic field, the oscillatory density of states at the Fermi energy that we will discuss later in the book can be directly measured. The term before the last one depends on the quantization energy of the electrons in the triangular potential of the structure. If the electron density increases, the potential becomes steeper, the quantization energy E_0 rises, and the width and center of mass of the ground state wave function decrease. This behavior will now be discussed within a quantum mechanical model for the system.

9.4 Fang–Howard variational approach

Calculating the quantization energy for electrons in a two-dimensional electron gas is a nontrivial problem of many-body physics without an analytic solution. The simplest approximation is to calculate the quantization energy in a fixed (i.e., density independent) triangular potential. Rather than following this very crude approximation we will discuss a variational approach minimizing the energy of the system by optimizing one variational parameter in the wave function. The hamiltonian of the electronic system is given by

$$H = \sum_i \left[-\frac{\hbar^2}{2m^\star}\Delta_i + U(z_i)\right] + \frac{1}{2}\sum_{i,j;j\neq i} G(\mathbf{r}_i, \mathbf{r}_j), \tag{9.2}$$

where $G(\mathbf{r}_i, \mathbf{r}_j)$ describes the Coulomb interaction between electrons in the electron gas. Compared to eq. (8.1) we consider the case of zero magnetic field (leading to spin degeneracy) and neglect image charge effects due to the gate electrode. The potential $U(z)$ contains the electrostatic

potential of the donors and the contribution of the band offset at the heterointerface.

Our first approximation will be to consider the height of the potential barrier at the heterointerface to be infinite. In this case, the wave function does not penetrate into the barrier, but rather vanishes at the interface. It is then sufficient to consider the problem for $z \geq 0$.

We will further treat the electron–electron interaction within the self-consistent Hartree approximation (cf., p. 109). In this case, the single-particle envelope wave functions $\psi_{n\mathbf{k}}(\mathbf{r})$ fulfill the Hartree equation

$$\left[-\frac{\hbar^2}{2m^\star}\Delta + V_{\mathrm{H}}(z)\right]\psi_{n\mathbf{k}}(\mathbf{r}) = E_n(\mathbf{k})\psi_{n\mathbf{k}}(\mathbf{r}), \tag{9.3}$$

and

$$V_{\mathrm{H}}(z) = \sum_{n\mathbf{k}} \int d^3r' G(\mathbf{r},\mathbf{r}') \left|\psi_{n\mathbf{k}}(\mathbf{r}')\right|^2$$

is the Hartree potential. The sum is taken over all states occupied at zero temperature. We can omit the $U(z_i)$ contribution in the hamiltonian (9.2) by requiring that the Hartree potential and its derivative vanish for $z \to \infty$. This will automatically generate the correct electric field at the heterointerface for a given sheet electron density n_{s}. As a result of the translational invariance of the problem in the x-y-plane, the Hartree potential depends only on z, and in the wave functions the parts depending on x- and y-coordinates in the plane can be separated leading to

$$\psi_{n\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{A}} e^{i\mathbf{k}\boldsymbol{\rho}} \varphi_n(z), \tag{9.4}$$

where $\mathbf{k}$ and $\boldsymbol{\rho}$ are an in-plane wave vector and a position vector, respectively. Using this *Ansatz* the energies are

$$E_n(\mathbf{k}) = E_n + \frac{\hbar^2 k^2}{2m^\star}$$

and the $\varphi_n(z)$ obey the one-dimensional Schrödinger equation

$$\left[-\frac{\hbar^2}{2m^\star}\frac{\partial^2}{\partial z^2} + V_{\mathrm{H}}(z)\right]\varphi_n(z) = E_n\varphi_n(z).$$

Although nowadays the problem can be solved self-consistently on a personal computer within seconds, we can improve our understanding of the system by using another approximative approach.

Following the idea of the Hartree approximation we use a product of single-particle wave functions of the form (9.4) to approximate the many-body wave function. However, unlike the usual Hartree-approximation, we do not calculate the wave functions self-consistently, but try to find a good approximation by using the so-called Fang–Howard variational wave function for the ground state in the z-direction

$$\varphi_0(z) = \sqrt{\frac{b^3}{2}} z e^{-bz/2}$$

which is normalized and accounts for our assumption that the wave function is zero at the heterointerface. The variational parameter b will now be determined such that the expectation value of the hamiltonian (9.2), normalized to the number of electrons, is minimized.

In analogy to eq. (8.10) this expectation value consists of two parts:

$$E = \langle \hat{T} \rangle + \frac{1}{2} \langle \hat{V}_{\mathrm{H}} \rangle,$$

i.e., the expectation value of the single-particle kinetic energy and that of the electron–electron interaction energy which is half the Hartree energy. Inserting the product wave function into the hamiltonian (9.2) we find the expectation value of the kinetic energy in the z-direction

$$\begin{aligned} \langle \hat{T} \rangle &= \frac{2}{A} \sum_{\mathbf{k}} \left(\frac{\hbar^2 k^2}{2m^\star} - \frac{\hbar^2}{2m^\star} \frac{b^3}{2} \int dz z e^{-bz/2} \frac{\partial^2}{\partial z^2} z e^{-bz/2} \right) \\ &= \left(\frac{1}{2} \frac{\pi \hbar^2}{m^\star} n_{\mathrm{s}} + \frac{\hbar^2 b^2}{8 m^\star} \right) n_{\mathrm{s}}. \end{aligned}$$

The first term represents the in-plane kinetic energy of the system, whereas the second part is the kinetic energy in the confinement direction.

In order to find the expectation value of the Hartree interaction $\langle V_{\mathrm{H}} \rangle$, we determine the electron density distribution

$$\rho(z) = -|e| n_{\mathrm{s}} \left| \varphi_0(z) \right|^2 = -|e| n_{\mathrm{s}} \frac{b^3}{2} z^2 e^{-bz}.$$

The Hartree potential is now obtained as a solution of Poisson's equation

$$\frac{\partial^2 V_{\mathrm{H}}(z)}{\partial z^2} = \frac{|e| \rho(z)}{\varepsilon \varepsilon_0}$$

with the boundary conditions $\partial V_{\mathrm{H}}(0)/\partial z = e^2 n_{\mathrm{s}}/\varepsilon\varepsilon_0$ and $V_{\mathrm{H}}(0) = 0$. The solution is found to be

$$V_{\mathrm{H}}(z) = \frac{e^2 n_{\mathrm{s}}}{2\varepsilon\varepsilon_0 b} \left\{ 6 - \left[(bz)^2 + 4bz + 6 \right] e^{-bz} \right\}.$$

Figure 9.3 shows the resulting potential. The function $V_{\mathrm{bare}}(z)$ represents the part of the potential created by the charged donors and the surface charges in the absence of the electron gas. The two-dimensional electron gas screens this potential in the z-direction making the electric field zero for $z \to \infty$.

With this result we obtain, for the expectation value of the Hartree potential,

$$\begin{aligned} \langle \hat{V}_{\mathrm{H}} \rangle &= \frac{e^2 n_{\mathrm{s}}^2}{2\varepsilon\varepsilon_0 b} \frac{b^3}{2} \int_0^\infty dz z^2 e^{-bz} \left\{ 6 - \left[(bz)^2 + 4bz + 6 \right] e^{-bz} \right\} \\ &= \frac{e^2 n_{\mathrm{s}}^2}{4\varepsilon\varepsilon_0 b} \int_0^\infty dx \left\{ 6x^2 e^{-x} - \left[x^4 + 4x^3 + 6x^2 \right] e^{-2x} \right\} \\ &= \frac{33 e^2 n_{\mathrm{s}}^2}{16 \varepsilon\varepsilon_0 b}. \end{aligned}$$

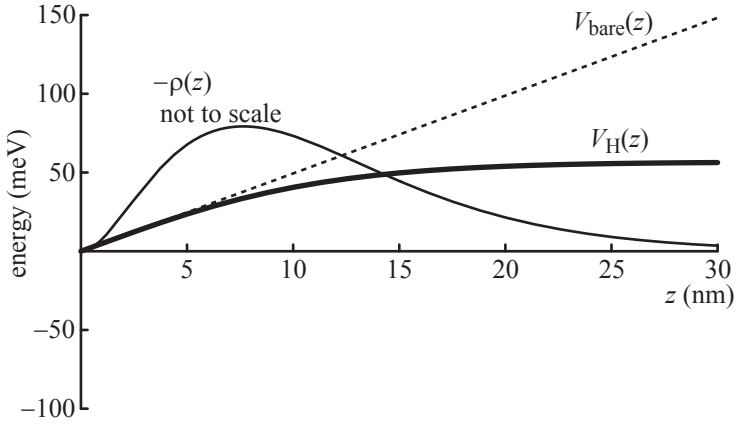


Fig. 9.3 Hartree potential normal to a heterointerface. The electron density distribution is plotted as well.

The total energy that is to be minimized is then

$$E = \left(\frac{\pi\hbar^2}{2m^\star}n_\mathrm{s} + \frac{\hbar^2 b^2}{8m^\star}\right) n_\mathrm{s} + \frac{33e^2 n_\mathrm{s}^2}{32\varepsilon\varepsilon_0 b}.$$

This expression is minimum if its derivative with respect to b is zero, resulting in

$$b = \left(\frac{33\pi}{2}n_\mathrm{s}a_\mathrm{B}^{\star 2}\right)^{1/3}\frac{1}{a_\mathrm{B}^\star}.$$

The result fulfills our expectation that the width of the wave function decreases with increasing electron density.

The corresponding ground state quantization energy is given by the expectation value of the single-particle hamiltonian (9.3) [see also eq. (8.11)]

$$E_0(n_\mathrm{s}) = \langle T\rangle + \langle V_\mathrm{H}\rangle = \frac{5}{4}E_\mathrm{Ry}^\star\left(\frac{33\pi}{2}n_\mathrm{s}a_\mathrm{B}^{\star 2}\right)^{2/3}.$$

The quantization energy in the z-direction increases with increasing electron density, in agreement with our intuition.

We briefly return to the capacitance between top gate and electron gas in eq. (9.1). There, a length scale appears containing the derivative of the quantization energy with respect to the electron density. With the above result this length scale becomes

$$\frac{\varepsilon\varepsilon_0}{e^2}\frac{dE_0(n_\mathrm{s})}{dn_\mathrm{s}} = \frac{55}{32}\left(\frac{33\pi}{2}n_\mathrm{s}a_\mathrm{B}^{\star 2}\right)^{-1/3}a_\mathrm{B}^\star.$$

This length scale decreases with increasing electron density. We compare it with the expectation value for the center of mass of the wave function

$$\langle z\rangle = \frac{b^3}{2}\int_0^\infty dz z^3 e^{-bz} = \frac{3}{b} = 3\left(\frac{33\pi}{2}n_\mathrm{s}a_\mathrm{B}^{\star 2}\right)^{-1/3}a_\mathrm{B}^\star$$

and find

$$\frac{\varepsilon\varepsilon_0}{e^2}\frac{dE_0(n_\mathrm{s})}{dn_\mathrm{s}} = \frac{55}{96}\langle z\rangle.$$

We see that this length scale describes the effective increase of the separation between metallic top gate and electron gas by the finite extent of the wave function. For a two-dimensional electron gas in GaAs with density $n_\mathrm{s} = 3 \times 10^{11}\,\mathrm{cm}^{-2}$ this additional separation is 6.9 nm and $\langle z \rangle = 12\,\mathrm{nm}$.

Influence of DX-centers on AlGaAs/GaAs heterostructures. In AlGaAs heterostructures, DX-centers have a profound influence on the tunability of the electron density. When designing a heterostructure, one has to take care that the energy of DX-states remains above the electrochemical potential of the two-dimensional electron gas. Otherwise the DX-centers can become electrically neutral and the tunability of the electron gas is almost entirely suppressed. Instead, the occupation of states in the doping plane is tuned. Even in structures where this effect has been accounted for in the design, at positive gate voltages the tunability of the electron density saturates, even before the Schottky barrier becomes transparent for electrons. This effect can, for example, be seen in Fig. 5.17.

9.5 Spatial potential fluctuations and the theory of screening

9.5.1 Spatial potential fluctuations

In the above discussion of the two-dimensional electron gas in a GaAs/AlGaAs heterostructure we have used the jellium model replacing the discrete dopant distribution by a smeared mean density. We will now discuss the influence of the neglected fluctuating part $C(\mathbf{r})\delta(z+s)$ of the dopant distribution on the potential landscape in the two-dimensional electron gas. Neglecting image charge effects, the corresponding electrostatic potential is

$$V(\mathbf{r}, z) = \frac{e^2}{4\pi\epsilon\epsilon_0} \int d^2r' \frac{C(\mathbf{r}')}{\sqrt{(\mathbf{r}-\mathbf{r}')^2 + (z+s)^2}}.$$

The in-plane average of this electrostatic potential is

$$\langle V(\mathbf{r}, z) \rangle = 0,$$

because $\langle C(\mathbf{r}) \rangle = 0$.

Frequently, the Fourier transform of this potential in the plane is of importance. It can be shown to be (see Appendix A.3):

$$V(\mathbf{q}, z) = \frac{e^2}{2\epsilon\epsilon_0} \frac{C(\mathbf{q})}{q} e^{-q|z+s|}.$$

We can see in this expression that the short-range contributions of the fluctuating potentials (large q) are exponentially damped. The larger

the spacer thickness s, the larger are the length scales on which the potential varies in the plane.

It turns out that the mean square fluctuation $\langle V^2(\mathbf{r},z)\rangle$ of the potential diverges in a sample of infinite size [see, e.g., Efros *et al.* 1993]. Therefore, the effect of screening by the electron gas is very important for the understanding of the magnitude of spatial potential fluctuations.

9.5.2 Linear static polarizability of the electron gas

General theory. A static potential $V_{\rm tot}(\mathbf{r},z)$ within the electron gas, as it is caused by the fluctuating part of the donor distribution, leads to the appearance of in-plane forces acting on the electrons. The result is a local modification of the electron density called an induced electron density $n_{\rm ind}(\mathbf{r},z)$.

The linear static polarizability $P(\mathbf{r},z;\mathbf{r}',z')$ of the electron gas is the linear response function relating the induced electron density with the static potential. It is a system property independent of the external potential. We write

$$n_{\rm ind}(\mathbf{r},z) = \int d^2r dz P(\mathbf{r},z;\mathbf{r}',z') V_{\rm tot}(\mathbf{r}',z'). \tag{9.5}$$

This nonlinear nonlocal polarizability was calculated by Hedin and Lundqvist in first order perturbation theory for electronic systems of arbitrary dimensionality (Hedin and Lundqvist, 1969). It is given by

$$P(\mathbf{r},\mathbf{r}') = 2\sum_{nm\mathbf{kq}} \frac{f(E_{n\mathbf{k}}) - f(E_{m\mathbf{k+q}})}{E_{n\mathbf{k}} - E_{m\mathbf{k+q}} - i0^+} \psi_{n\mathbf{k}}(\mathbf{r})\psi^\star_{m\mathbf{k+q}}(\mathbf{r})\psi^\star_{n\mathbf{k}}(\mathbf{r}')\psi_{m\mathbf{k+q}}(\mathbf{r}').$$

Here $f(E)$ is the Fermi–Dirac distribution function, and the wave functions $\psi_{n\mathbf{k}}(\mathbf{r})$ are the (envelope) wave functions of the state with quantum numbers $n\mathbf{k}$ in the unperturbed system. In a two-dimensional electron gas the unperturbed wave functions are given by eq. (9.4). Including them into the above expression gives

$$P(\mathbf{r}-\mathbf{r}';z,z') = \frac{1}{4\pi^2}\int d^2q\, e^{-i\mathbf{q}(\mathbf{r}-\mathbf{r}')} P(q;z,z'),$$

where

$$P(q;z,z') = \sum_{nm} \Pi_{nm}(q,\mu,T)\varphi_n(z)\varphi^\star_m(z)\varphi^\star_n(z')\varphi_m(z')$$

with

$$\Pi_{nm}(q,\mu,T) = \frac{1}{2\pi^2}\int d^2k\, \frac{f(E_{n\mathbf{k}}) - f(E_{m\mathbf{k+q}})}{E_{n\mathbf{k}} - E_{m\mathbf{k+q}} - i0^+}.$$

As a consequence the integral on the right-hand side of eq. (9.5) is a two-dimensional convolution integral and the two-dimensional Fourier transform (Fourier–Bessel transform, see Appendix A.3) of the induced electron density in the plane is

$$n_{\rm ind}(\mathbf{q},z) = \int dz' P(q;z,z') V(\mathbf{q},z'). \tag{9.6}$$

The polarization function $\Pi_{nm}(q,\mu,T)$ appearing in the polarizability can be calculated exactly for two-dimensional systems with parabolic dispersion at zero temperature. The result is

$$\begin{aligned}\Pi_{nm}(q,E_F,0) &= \frac{m^\star}{\pi\hbar^2}\left[\frac{k_F^{(n)}}{q}I\left(\frac{E_n-E_m}{2qk_F^{(n)}}-\frac{q}{2k_F^{(n)}}\right)\right.\\ &\left.-\frac{k_F^{(m)}}{q}I\left(\frac{E_n-E_m}{2qk_F^{(m)}}+\frac{q}{2k_F^{(m)}}\right)\right]\end{aligned}$$

with

$$I(x)=\begin{cases} x+i\sqrt{1-x^2} & \text{f. } |x|<1 \\ x-\mathrm{sgn}(x)\sqrt{x^2-1} & \text{f. } |x|>1 \end{cases}.$$

Polarization for systems with only one occupied subband. Assume that only one quantized subband in the z-direction is occupied (quantum limit) in a system with parabolic dispersion. The wave function of this subband is denoted with $\varphi_0(z)$. In this case the polarizability is well described by

$$P(\mathbf{r},z;\mathbf{r}',z')=\frac{1}{(2\pi)^2}\int d^2q\Pi_{00}(q,E_\mathrm{F},T)e^{-i\mathbf{q}(\mathbf{r}-\mathbf{r}')}\left|\varphi_0(z)\right|^2\left|\varphi_0(z')\right|^2, \tag{9.7}$$

and therefore

$$P(q;z,z')=\Pi_{00}(q,E_\mathrm{F},T)\left|\varphi_0(z)\right|^2\left|\varphi_0(z')\right|^2.$$

At temperature $T=0$ we have

$$\Pi_{00}(q,E_F,0)=-\frac{m^\star}{\pi\hbar^2}\begin{cases}1 & \text{for } q\le 2k_F \\ 1-\sqrt{1-\frac{4k_F^2}{q^2}} & \text{for } q>2k_F\end{cases}.$$

Note here that $m^\star/\pi\hbar^2=\mathcal{D}_{2\mathrm{D}}$ is the two-dimensional density of states. At finite temperatures the polarization function can be calculated from the zero temperature result using the formula of Maldague (Maldague, 1978):

$$\Pi(q,T,\mu_\mathrm{ch})=\int_0^\infty\frac{\Pi(q,T=0,\mu')d\mu'}{4kT\cosh^2[(\mu_\mathrm{ch}-\mu')/2kT]}. \tag{9.8}$$

Figure 9.4 shows the polarization function for different temperatures. At $T=0$ the function is constant for $q<2k_F$. At $q=2k_F$ there is a kink known as the Kohn anomaly (or Kohn singularity). It is a singularity, because the function does not have a well-defined derivative at $q=2k_\mathrm{F}$. For larger wave vectors, the function decreases monotonically. At finite temperatures the Kohn anomaly is smeared out. Most importantly, the temperature dependence of the polarization function is most pronounced around $q=2k_F$.

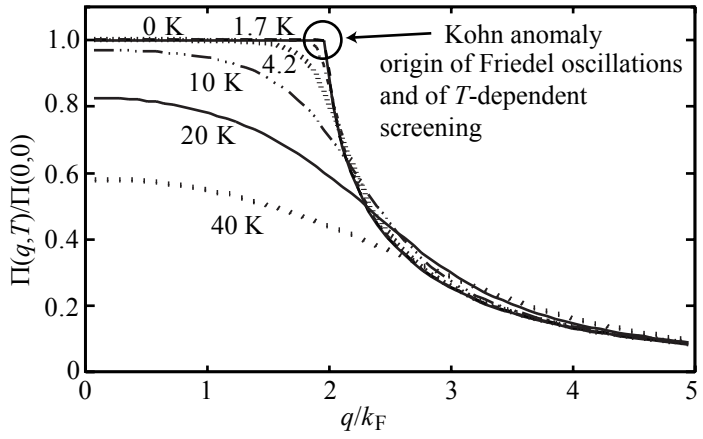


Fig. 9.4 Polarization function $\Pi_{00}(q, \mu, T)$ as a function of q for a number of temperatures.

Thomas–Fermi approximation. The Fourier transform of the polarization function, $\Pi_{00}(q, E_\mathrm{F}, T)$, has been found to be independent of q for $q \leq 2k_\mathrm{F}$. For the screening of long-range fluctuating potentials it is a good approximation to assume that the function is q-independent for all q. This approximation is identical to the Thomas–Fermi approximation. In the two-dimensional quantum limit

$$P(\mathbf{r}, z; \mathbf{r}', z') = -\mathcal{D}_{2\mathrm{D}} \delta(\mathbf{r} - \mathbf{r}') \left|\varphi_0(z)\right|^2 \left|\varphi_0(z')\right|^2,$$

i.e., the polarizability becomes a local function and the induced density is given by

$$n_\mathrm{ind}(\mathbf{r}, z) = -\mathcal{D}_{2\mathrm{D}} \left|\varphi_0(z)\right|^2 \int dz' \left|\varphi_0(z')\right|^2 V_\mathrm{tot}(\mathbf{r}, z').$$

The remaining integral is the expectation value of the potential normal to the plane of the electron gas which we will denote with $\langle V_\mathrm{tot}(\mathbf{r})\rangle$. The distribution of the induced density in the z-direction follows the distribution of the ground state wave function. In the plane the electrons fill the density of states at each position $\mathbf{r}$ up to the energy $\langle V_\mathrm{tot}(\mathbf{r})\rangle$. This is the characteristic feature of the Thomas–Fermi approximation. It can be successfully used, if the dominant Fourier components of $V_\mathrm{tot}(\mathbf{r})$ arise for $q \ll 2k_\mathrm{F}$, i.e., in the case of long-range fluctuating potentials arising as a result of a sufficiently thick spacer layer between dopants and electron gas. Quantitatively this means

$$2k_\mathrm{F} s \gg 1.$$

9.5.3 Linear screening

External, induced and total potential. If an external potential acts on a two-dimensional electron gas, the induced electron density will itself give rise to a potential. The total potential in which the electrons move can be calculated self-consistently. This method is, in this context,

also called random phase approximation (RPA). Within this method the total potential is the sum of the external and the induced Hartree potential:

$$V_{\rm tot}(\mathbf{r}, z) = V_{\rm ext}(\mathbf{r}, z) + V_{\rm ind}(\mathbf{r}, z).$$

In Fourier space this relation reads

$$V_{\rm tot}(\mathbf{q}, z) = V_{\rm ext}(\mathbf{q}, z) + V_{\rm ind}(\mathbf{q}, z).$$

Induced potential from the induced density. While the external potential is given as created by the distribution of dopants, the induced potential is obtained as the solution of Poisson's equation [cf., eq. (7.4)], i.e.,

$$V_{\rm ind}(\mathbf{r}) = e^2 \int d^3r'\, G(\mathbf{r}, \mathbf{r}') n_{\rm ind}(\mathbf{r}'),$$

where $G(\mathbf{r}, \mathbf{r}')$ is Green's function (cf., p. 96). Neglecting image charge effects due to the presence of gate electrodes the well-known solution reads

$$G(\mathbf{r}, z; \mathbf{r}', z') = \frac{1}{4\pi\epsilon\epsilon_0} \frac{1}{\sqrt{(\mathbf{r} - \mathbf{r}')^2 + (z - z')^2}} \tag{9.9}$$

with the in-plane Fourier transform

$$V_{\rm ind}(\mathbf{q}, z) = e^2 \int dz'\, G(q; z, z') n_{\rm ind}(\mathbf{q}, z'). \tag{9.10}$$

The Fourier transform $G(q; z, z')$ of Green's function (see Appendix A.3) is given by

$$G(q; z, z') = \frac{1}{2\epsilon\epsilon_0 q} e^{-q|z-z'|} \text{ for } q \neq 0. \tag{9.11}$$

Induced potential and external potential. The induced density is related to the total potential via the polarizability of the two-dimensional electron gas. As a result

$$V_{\rm ind}(\mathbf{r}, z) = e^2 \int d^2r' dz' G(\mathbf{r}, z; \mathbf{r}', z') \int d^2r'' dz'' P(\mathbf{r}', z'; \mathbf{r}'', z'') V_{\rm tot}(\mathbf{r}'', z'').$$

This leads to the self-consistent equation for the total potential

$$V_{\rm tot}(\mathbf{r}, z) = V_{\rm ext}(\mathbf{r}, z) + \int d^2r' dz' G(\mathbf{r}, z; \mathbf{r}', z') \int d^2r'' dz'' P(\mathbf{r}', z'; \mathbf{r}'', z'') V_{\rm tot}(\mathbf{r}'', z'').$$

We express this result as

$$V_{\rm ext}(\mathbf{r}, z) = \int d^2r' dz' \epsilon(\mathbf{r}, z; \mathbf{r}', z') V_{\rm tot}(\mathbf{r}', z'),$$

defining Lindhard's nonlocal dielectric function

$$\epsilon(\mathbf{r}, z; \mathbf{r}', z') := \delta(\mathbf{r}-\mathbf{r}')\delta(z-z') - \int d^2r'' dz'' G(\mathbf{r}, z; \mathbf{r}'', z'') P(\mathbf{r}'', z''; \mathbf{r}', z').$$

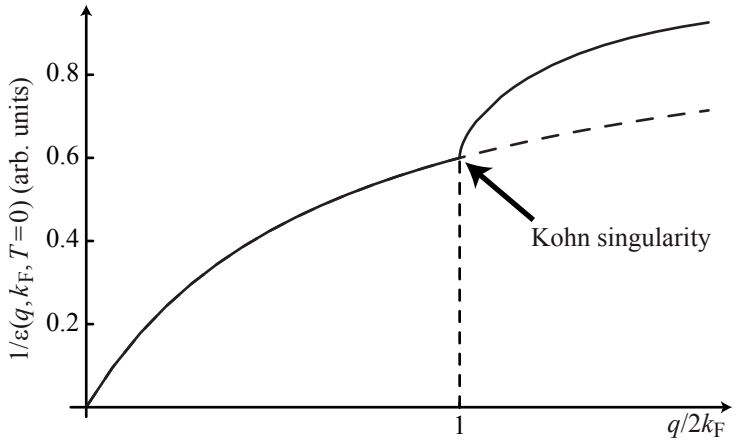


Fig. 9.5 Inverse dielectric function of a two-dimensional electron gas in the quantum limit at zero temperature. The dashed line is the dielectric function in Thomas–Fermi approximation.

Screening with only one occupied subband: quantum limit. If only one subband is occupied the expression for Lindhard's nonlocal dielectric function simplifies considerably. Using eqs. (9.9) and (9.7) we obtain an expression for the relation between the expectation values of the external and the total potential in Fourier space:

$$\underbrace{\left[1 - \frac{e^2}{2\epsilon\epsilon_0 q}\Pi_{00}(q, E_\mathrm{F}, T)F(q)\right]}_{:=\epsilon(q,E_\mathrm{F},T)} \langle V_\mathrm{tot}(\mathbf{q})\rangle = \langle V_\mathrm{ext}(\mathbf{q})\rangle$$

Here we have introduced the form factor

$$F(q) = \int dz'dz\, |\varphi_0(z)|^2\, e^{-q|z-z'|}\, |\varphi_0(z')|^2$$

and Lindhard's dielectric function $\epsilon(q, E_\mathrm{F}, T)$ in Fourier space. We obtain the relation

$$\langle V_\mathrm{tot}(\mathbf{q})\rangle = \frac{\langle V_\mathrm{ext}(\mathbf{q})\rangle}{\epsilon(q, E_\mathrm{F}, T)}. \tag{9.12}$$

Figure 9.5 shows the inverse dielectric function $\epsilon^{-1}(q, \mu, T)$ schematically. The effect of screening is the suppression of Fourier components with small q, i.e., the long-range parts of the potential in real space.

Thomas–Fermi approximation. The Thomas–Fermi approximation introduced above leads to a further significant simplification of the dielectric function. We obtain

$$\epsilon_\mathrm{TF}(\mathbf{q}) = 1 + \frac{2}{qa_\mathrm{B}^\star}$$

if the $q \to 0$ limit of the form factor is used. In Fig. 9.5, the dielectric function in this approximation is shown as the dashed line. The quantity

$$\lambda_\mathrm{TF} = \pi a_\mathrm{B}^\star$$

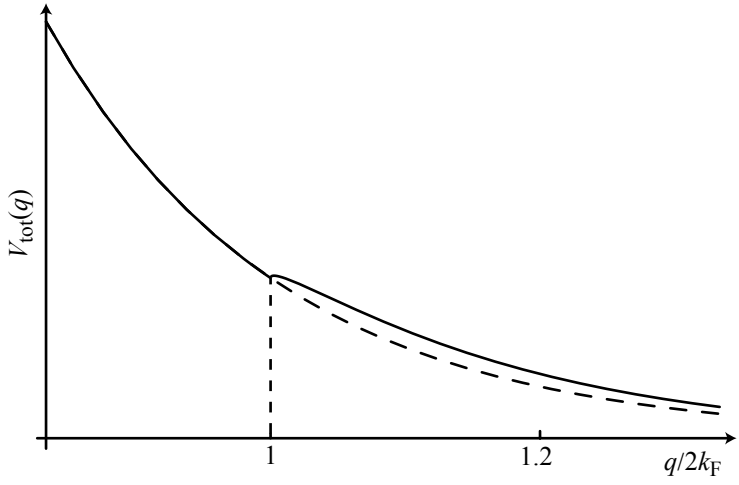


Fig. 9.6 Screened potential of a charged dopant in Fourier space at temperature $T = 0$. Only a region around $q = 2k_{\rm F}$ is plotted. The dashed line corresponds to the Thomas–Fermi approximation.

is called the *Thomas–Fermi screening length*. For two-dimensional electron gases it is independent of the electron density. We define the Thomas–Fermi wave vector as

$$q_{\rm TF} = \frac{2}{a_{\rm B}^\star} = \frac{2\pi}{\lambda_{\rm TF}}.$$

9.5.4 Screening a single point charge

As an example of the application of screening theory in two dimensions, we consider screening of a point charge Ze placed at $(x, y, z) = (0, 0, d)$, i.e., at a distance d from the plane of a two-dimensional electron gas. Neglecting image charge effects, the in-plane potential is given by

$$V_{\rm ext}(r) = -\frac{Ze^2}{4\pi\epsilon\epsilon_0}\frac{1}{\sqrt{r^2 - d^2}},$$

with the Fourier transform (see appendix A.3)

$$V_{\rm ext}(q) = -\frac{Ze^2}{2\epsilon\epsilon_0}\frac{1}{q}e^{-qd}.$$

The Fourier transform of the total potential is, at $T = 0$, given by

$$V_{\rm tot}(q) = -\frac{Ze^2}{2\epsilon\epsilon_0}\frac{e^{-qd}}{q + q_{\rm TF}g(q)},$$

where $g(q) = \Pi_{00}(q, E_{\rm F}, 0)/\Pi_{00}(0, E_{\rm F}, 0)$. Figure 9.6 shows this potential around $q = 2k_{\rm F}$. It can be seen how the Kohn singularity of the dielectric function is transferred to the total potential.

Long range part of the potential. In order to understand the behavior of the potential in real space we concentrate first on the long range part. To this end we expand the potential around $q = 0$ up to first order.

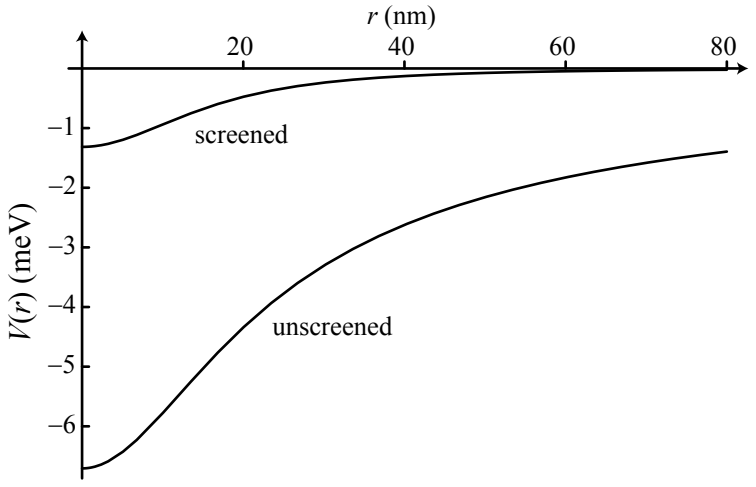


Fig. 9.7 Coulomb potential of a unity point charge at a distance of 17 nm from a two-dimensional electron gas in a Ga[Al]As heterostructure with and without screening in Thomas–Fermi approximation ($T = 0$).

In this approximation the Kohn singularity does not play a role, and we are essentially working in the Thomas–Fermi approximation. We obtain

$$V_{\rm tot}(q) = -\frac{Ze^2}{2\epsilon\epsilon_0}\frac{1}{q_{\rm TF}}\left[1-(1+q_{\rm TF}d)\frac{q}{q_{\rm TF}}\right] + \mathcal{O}\left(q^2\right).$$

Transformation back to real space leads to (Stern, 1967)

$$V_{\rm tot}(r) = -\frac{Ze^2 q_{\rm TF}}{4\pi\epsilon\epsilon_0}\frac{1+q_{\rm TF}d}{(q_{\rm TF}r)^3} + \mathcal{O}\left(r^{-5}\right) \quad \text{f. } r\to\infty. \tag{9.13}$$

For large distances from the charged donor, the potential decays as r^{-3}, i.e., faster than the bare Coulomb potential. In addition, we calculate the amplitude of the screened potential in real space at $r = 0$ setting $g(q) = 1$ (Thomas–Fermi approximation) for simplicity. We obtain

$$\begin{aligned} V_{\rm tot}(r=0) &= -\frac{Ze^2}{4\pi\epsilon\epsilon_0}\int_0^\infty dq\; qJ_0(0)\frac{e^{-qd}}{q+q_{\rm TF}} \\ &= -\frac{Ze^2}{4\pi\epsilon\epsilon_0 d}\left[1-q_{\rm TF}de^{q_{\rm TF}d}\Gamma(0,q_{\rm TF}d)\right]. \end{aligned}$$

Here $J_0(x)$ is a Bessel function and $\Gamma(a,x)$ is the incomplete gamma function. The amplitude of the potential modulation at $r = 0$ decreases with increasing $q_{\rm TF}d$ and tends to zero for $d\to\infty$. If d is about one third of $a_{\rm B}^\star$, the amplitude of the screened potential is already about half of the unscreened potential. Figure 9.7 shows the influence of screening in Thomas–Fermi approximation on the potential of a positive unit charge located at a distance of 17 nm from a two-dimensional electron gas in a Ga[Al]As heterostructure.

Friedel oscillations. Within the Thomas–Fermi approximation, valid only for $2k_{\rm F}d \ll 1$, we have omitted effects originating from large q

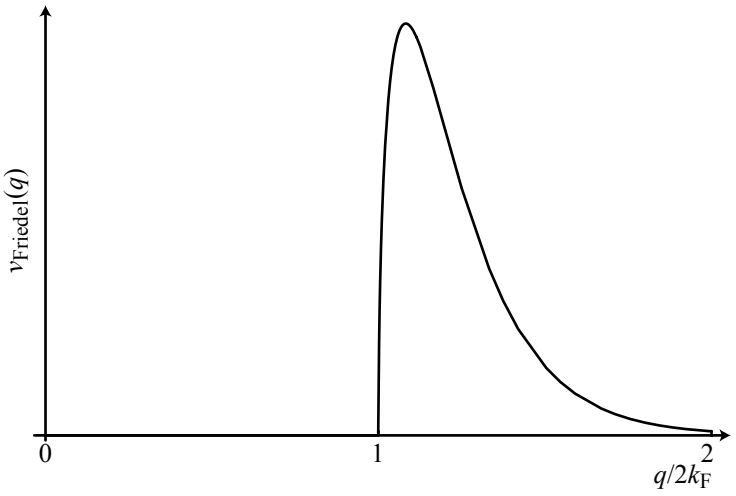


Fig. 9.8 Potential $v_{\mathrm{Friedel}}(q)$ (at $T = 0$).

and Kohn's singularity. The full expression for the screened Coulomb potential is

$$V_{\mathrm{tot}}(r) = -\frac{Ze^2}{4\pi\epsilon\epsilon_0}\int_0^\infty dq\; qJ_0(qr)\underbrace{\frac{e^{-qd}}{q + q_{\mathrm{TF}}g(q)}}_{:=v_{\mathrm{tot}}(q)}.$$

For further discussion we split the screened potential into a contribution V_{TF} obtained within the Thomas–Fermi approximation, and a contribution V_{Friedel}. In Fourier space this leads to

$$v_{\mathrm{tot}}(q) = \frac{v_{\mathrm{ext}}}{\epsilon_{\mathrm{TF}}(q)} + v_{\mathrm{Friedel}}(q),$$

where

$$v_{\mathrm{Friedel}}(q) = q_{\mathrm{TF}}e^{-qd}\frac{1 - g(q)}{[q + q_{\mathrm{TF}}g(q)](q + q_{\mathrm{TF}})}.$$

This expression is zero for $q < 2k_{\mathrm{F}}$. At $q = 2k_{\mathrm{F}}$ it has the singularity as shown in Fig. 9.8.

We investigate the asymptotic behavior of $v_{\mathrm{Friedel}}(r)$ for large r ($2k_{\mathrm{F}}r \gg 1$). In this limit we can use the limiting expression of the Bessel function in the Fourier integral

$$J_0(qr) = \sqrt{\frac{2}{\pi qr}}\cos\left(qr - \frac{\pi}{4}\right) + \mathcal{O}\left(\frac{1}{qr}\right),$$

leading to a one-dimensional Fourier transform. An asymptotic expression for the screened potential at large distances is obtained using a theorem from Fourier transformation theory (see Lighthill, 1964, Chapter 4.3, Theorem 19). In order to apply this theorem we have to investigate the singularity further. Near $q = 2k_{\mathrm{F}}$ the function $\sqrt{q}v_{\mathrm{Friedel}}(q)$

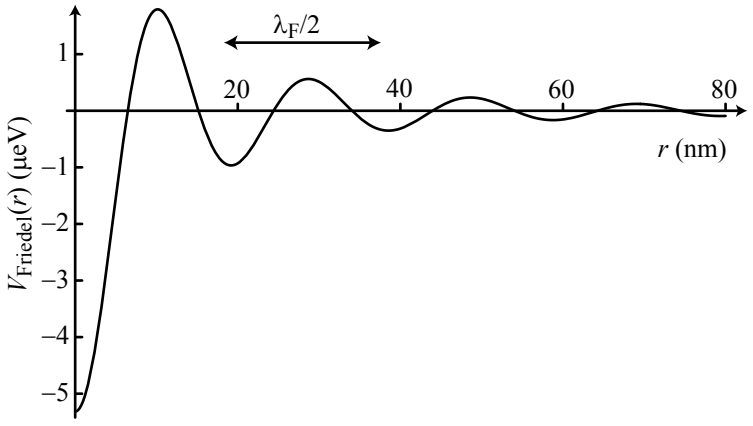


Fig. 9.9 Friedel oscillations in the screened potential of a point charge 17 nm remote from the two-dimensional electron gas in a Ga[Al]As heterostructure, calculated for the temperature $T = 0$.

behaves as

$$\begin{aligned}\sqrt{q}v_{\mathrm{Friedel}}(q) &= q_{\mathrm{TF}}\frac{\sqrt{2k_{\mathrm{F}}}e^{-2k_{\mathrm{F}}d}}{(2k_{\mathrm{F}}+q_{\mathrm{TF}})^2}\Theta(q-2k_{\mathrm{F}}) \\ &\quad \left\{\sqrt{\frac{q-2k_{\mathrm{F}}}{k_{\mathrm{F}}}}+\frac{q_{\mathrm{TF}}}{2k_{\mathrm{F}}+q_{\mathrm{TF}}}\frac{q-2k_{\mathrm{F}}}{k_{\mathrm{F}}}+\cdots\right\}.\end{aligned}$$

Transforming the first two terms back to real space leads to

$$\begin{aligned}V_{\mathrm{tot}}(r) &= -\frac{Ze^2q_{\mathrm{TF}}}{4\pi\epsilon\epsilon_0}\frac{4k_F^2}{(2k_F+q_{\mathrm{TF}})^2}e^{-2k_Fd} \\ &\quad \left[\frac{\sin 2k_Fr}{4k_F^2r^2}+\frac{\sqrt{8}q_{\mathrm{TF}}\cos(2k_Fr-\pi/4)}{\sqrt{\pi}(2k_F+q_{\mathrm{TF}})(2k_Fr)^{5/2}}+\cdots\right] \quad (9.14)\end{aligned}$$

Kohn's singularity in the dielectric function leads to an oscillatory behavior of the screened potential in real space with a wavelength of $\lambda_{\mathrm{F}}/2$, called Friedel oscillations. An example is shown in Figs. 9.9 and 9.10 for a Ga[Al]As heterostructure, with the point charge e 17 nm remote from the electron gas. For the form factor, $F(q) = 1$ was assumed corresponding to an ideal two-dimensional electron gas without extent in the z-direction. For increasing separation d between charged donor and electron gas, the amplitude of the Friedel oscillations decreases exponentially. For $d = 17$ nm, for example, the characteristic energy scale of the oscillations is almost three orders of magnitude smaller than the amplitude of the Thomas–Fermi potential contribution at $r = 0$.

Experimentally, Friedel oscillations can be measured, for example, on metallic surfaces on which a two-dimensional electron gas forms. A suitable system is an atomically flat Cu(111) surface. Figure 9.11 shows standing waves on such a surface with two lattice perturbations (Crommie *et al.*, 1993).

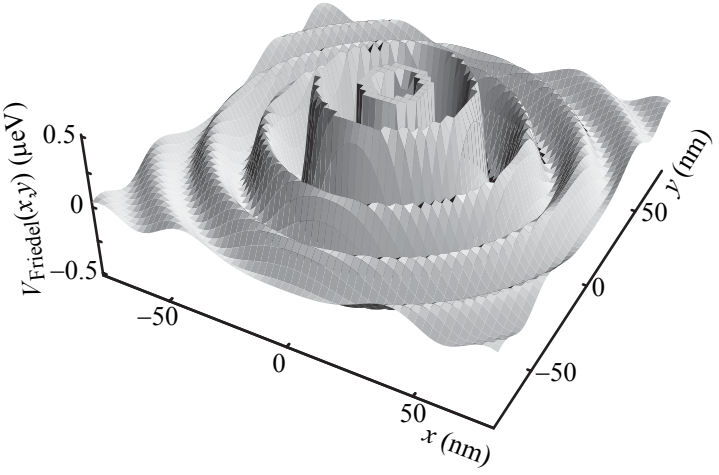


Fig. 9.10 Friedel oscillations in the plane. Parameters as for Fig. 9.9.

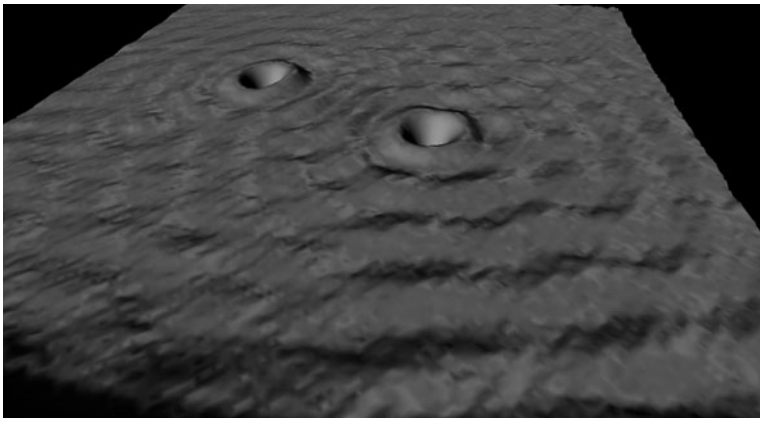


Fig. 9.11 Standing waves observed on a Cu(111) surface as observed with scanning tunneling microscope spectroscopy. Two point defects on the surface cause Friedel oscillations in the electron density that have here been mapped with real space and energy resolution (Crommie *et al.*, 1993).

9.5.5 Mean amplitude of potential fluctuations

Now we return to the discussion of the screened potential of a particular random distribution of donors. Within the Thomas–Fermi approximation we obtain

$$V_{\text{tot}}(\mathbf{r},z) = \frac{e^2}{2\epsilon\epsilon_0}\int \frac{d^2q}{(2\pi)^2}\,\frac{C(\mathbf{q})}{q+q_{\text{TF}}}e^{-q|z+s|}e^{i\mathbf{q}\mathbf{r}}.$$

The average fluctuation amplitude is therefore

$$\left\langle V_{\text{tot}}^2(\mathbf{r},z)\right\rangle = \left(\frac{e^2}{8\pi^2\epsilon\epsilon_0}\right)^2\int d^2q_1\int d^2q_2\frac{\langle C(\mathbf{q}_1)C(\mathbf{q}_2)\rangle}{(q_1+q_{\text{TF}})(q_2+q_{\text{TF}})}e^{-(q_1+q_2)|z+s|}e^{i(\mathbf{q}_1+\mathbf{q}_2)\mathbf{r}}. \tag{9.15}$$

Here, the brackets $\langle\ldots\rangle$ denote the average over a large number of different donor distributions. For calculating the correlator $\langle C(\mathbf{q}_1)C(\mathbf{q}_2)\rangle$

we transform to real space

$$\langle C(\mathbf{q}_1)C(\mathbf{q}_2)\rangle = \int d^2r_1 \int d^2r_2 \; \langle C(\mathbf{r}_1)C(\mathbf{r}_2)\rangle \, e^{-i\mathbf{q}_1\mathbf{r}_1} e^{-i\mathbf{q}_2\mathbf{r}_2}.$$

If the donors are randomly placed in the doping plane, there is no correlation between their positions and we have

$$\langle C(\mathbf{r}_1)C(\mathbf{r}_2)\rangle = N_\mathrm{d}\delta(\mathbf{r}_1 - \mathbf{r}_2).$$

As a result

$$\langle C(\mathbf{q}_1)C(\mathbf{q}_2)\rangle = (2\pi)^2 N_\mathrm{d}\delta(\mathbf{q}_1 + \mathbf{q}_2)$$

and the mean squared amplitude of the fluctuating potential is given by

$$\left\langle V_\mathrm{tot}^2(\mathbf{r}, z)\right\rangle = \left(\frac{e^2}{4\pi\epsilon\epsilon_0}\right)^2 N_\mathrm{d} \int d^2q_1 \frac{1}{(q_1 + q_\mathrm{TF})^2} e^{-2q_1|z+s|}.$$

The integration can be performed analytically giving

$$\left\langle V_\mathrm{tot}^2(\mathbf{r}, z)\right\rangle = 2\pi \left(\frac{e^2}{4\pi\epsilon\epsilon_0}\right)^2 N_\mathrm{d} f(2q_\mathrm{TF}|z+s|),$$

with

$$f(x) = e^x \, (1 + x) \, \Gamma(0, x) - 1.$$

Here, $\Gamma(a, x)$ is the incomplete gamma function. Figure 9.12 shows the function $f(x)$. The mean amplitude of the potential roughness depends only on the parameter $2q_\mathrm{TF}|z+s|$. *The mean squared fluctuation amplitude is, in particular, independent of the electron density. It is therefore often said that screening in two dimensions is independent of the density.* This statement is only true in the context of linear screening and within the Thomas–Fermi approximation. It is essentially due to the constant value of the Thomas–Fermi wave vector q_TF. We also note here that density independent screening does not imply that the scattering rates

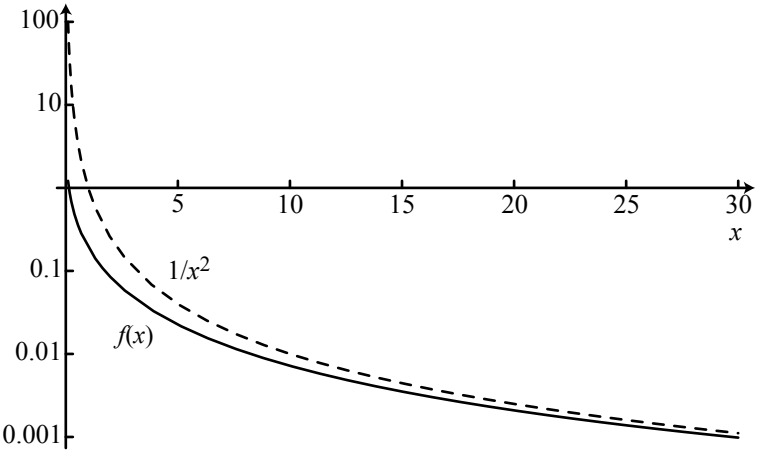


Fig. 9.12 The function $f(x)$ compared to the approximation $1/x^2$.

of electrons (or the electron mobilities) are independent of the electron densities.

In the limit $2q_{\mathrm{TF}}|z+s| \gg 1$ we have $f(x) \sim 1/x^2$, i.e., with increasing separation of the doping plane from the electron gas the amplitude of the spatially fluctuating potential decreases significantly. As a result, a lower scattering rate and a higher conductivity can be expected.

9.5.6 Nonlinear screening

Within the linear screening approximation, the divergence of $\left\langle V_{\mathrm{ext}}^2(\mathbf{r},z)\right\rangle$ arising in the unscreened case is cured and a finite value is reached for $\left\langle V_{\mathrm{tot}}^2(\mathbf{r},z)\right\rangle$. The linear screening approximation is useful, as long as

$$\sqrt{\langle V_{\mathrm{tot}}^2(\mathbf{r},z)\rangle} \ll E_{\mathrm{F}}.$$

Using the above expression for the mean amplitude of the fluctuating potential and the approximation $f(x) \sim 1/x^2$ we obtain the condition

$$R_{\mathrm{c}} \ll \frac{2|z+s|}{\sqrt{2\pi}},$$

where the quantity

$$R_{\mathrm{c}} = \frac{\sqrt{N_{\mathrm{d}}}}{n}$$

can be called the nonlinear screening length. Fluctuations with a wavelength larger than R_{c} are well screened and can be described within linear screening theory. Fluctuations on shorter length scales, however, are badly screened. The theory of nonlinear screening in two-dimensional electron gases has been developed by Efros (Efros, 1989).

At large electron densities $R_{\mathrm{c}} \ll |z+s|$ and all significant components of the unscreened potential are screened within the linear theory. If the density of the electron gas is reduced, the nonlinear screening length increases and eventually reaches the range $R_{\mathrm{c}} \sim |z+s|$. In this range the density distribution in the plane becomes very inhomogeneous on the length scale $|z+s|$. At even smaller densities, strong density fluctuations arise with wavelengths between $|z+s|$ and R_{s}.

Beyond the validity of linear screening, the idea behind the Thomas–Fermi approximation has still been used. The induced electron density is obtained by filling the density of states locally from the energy of the fluctuating potential up to the Fermi energy. Note that, in the case of a repulsive potential, the total electron density can never be smaller than zero. For example, if the electron density is more and more reduced by the application of a negative gate voltage, more and more potential hills will appear above the Fermi energy. A further reduction of the electron density leads to the localization of electrons in puddles enclosed by high potential barriers. The ability of electrons to percolate through such a system depends strongly on the exact potential landscape and electron density. A critical electron density n_{c} exists, at which percolation ceases because there is no connected domain of finite electron

density extending throughout a given sample. It has been shown using percolation theoretical considerations (Efros, 1989) that

$$n_{\mathrm{c}} = \beta \frac{\sqrt{N_{\mathrm{d}}}}{|z+s|},$$

where $\beta = 0.11$ is a numerical parameter. For a donor density $N_{\mathrm{d}} = 10^{12}\,\mathrm{cm}^{-2}$ and a separation $|z+s| = 50\,\mathrm{nm}$ between electron gas and doping plane, the critical density is $n_{\mathrm{c}} = 2.2 \times 10^{10}\,\mathrm{cm}^{-2}$.

9.6 Spin–orbit interaction

In the previous considerations we have assumed that the dispersion relation for the electrons in the plane is parabolic, with a curvature given by the effective mass, and independent of the electron spin. It turns out, that spin–orbit interaction can modify this simple picture, depending on the material in which the two-dimensional electron gas is realized.

We have already discussed the two-fold effect of spin–orbit interaction on the band structure in three-dimensional semiconductors (see pages 30 and 44). On the one hand there is the spin–orbit split-off band in all diamond or zinc-blende semiconductors, and, on the other, bulk inversion asymmetry leads to the Dresselhaus contribution to the spin–orbit interaction which changes the symmetry of the dispersion relation in a given energy band.

In two-dimensional systems we can distinguish two spin–orbit-related influences on the dispersion within a given energy band. On the one hand, as in three dimensions, the lack of inversion symmetry of the crystal lattice (bulk inversion asymmetry, giving rise to the Dresselhaus term in the hamiltonian), and on the other hand, the epitaxially grown structure, can create a confinement potential without spatial inversion symmetry (Bychkov and Rashba, 1984*a*; Bychkov and Rashba, 1984*b*). In this case we talk about structure inversion asymmetry (SIA, giving rise to the Rashba term in the hamiltonian).

The hamiltonian for two-dimensional electron gases in zinc blende heterostructures grown on [100] substrates is in lowest order in $\mathbf{k}$ given by

$$H = H_0 + \alpha_{\mathrm{R}}(\sigma_x k_y - \sigma_y k_x) + \beta_{\mathrm{D}}(\sigma_x k_x - \sigma_y k_y).$$

Here, σ_x and σ_y are the two components of Pauli's spin matrices in the plane of the two-dimensional electron gas. The first term H_0 describes the energy of the electrons in the absence of spin–orbit interaction. The second term is the Rashba term caused by structure inversion asymmetry. The third term is the Dresselhaus term describing the lack of inversion symmetry of the crystal structure.

The Dresselhaus term can be derived from the bulk Dresselhaus hamiltonian [eq. (3.29)] by taking the expectation value in the z-direction and keeping only terms linear in $\mathbf{k}$. The Dresselhaus coefficient β_{D} is given by the band structure parameters of the material and by the thickness

of the electron gas in the growth direction:

$$\beta_{\mathrm{D}} = \beta\langle k_z^2\rangle.$$

Here, $\langle k_z^2\rangle$ is the expectation value of the square of the wave vector in the confinement direction, i.e., of the order of $(\pi/W)^2$, with W being the width of the potential well. Narrow wells (small W) result in large values, wide wells (large W) smaller ones.

The form of the Rashba term, and its coefficient α_{R}, is related to the spin–orbit interaction hamiltonian (3.16) which contains an electric field. As discussed in Winkler 2003, Chapter. 6.3.2, a net electric field in the z-direction (growth direction) normal to the plane of the two-dimensional electron gas can arise as a result of the joint action of structure inversion asymmetry and admixture of valence band states to conduction band states. Its strength can be modified by the application of an external electric field E_z normal to the plane of the two-dimensional electron gas and therefore

$$\alpha_{\mathrm{R}} = \alpha\langle E_z\rangle,$$

with α being a material-specific constant and $\langle E_z\rangle$ being an electric field averaged in the z-direction. A detailed discussion on how this averaging has to be performed can again be found in Winkler 2003, Chapter. 6.3.2. Typical values for electric fields in heterostructures are a few mV/Å.

Values for α and β for a few common materials are given in Table 9.1. The table shows that in InSb the effects of spin–orbit interaction are extraordinarily high. Also in InAs, the Rashba effect is even more important than in other materials. In GaAs, AlAs, CdTe, and ZnSe the Rashba effect is small compared to the Dresselhaus effect. InAs is a suitable material for the investigation of the Rashba effect. In InSb both effects are comparable.

The action of the spin–orbit interaction on the spin of an electron in a particular state $\mathbf{k}$ can be described by an effective $\mathbf{k}$-dependent spin–orbit-induced magnetic field $\mathbf{B}_{\mathrm{SO}}(\mathbf{k})$ which allows the spin–orbit term in the hamiltonian to be written in the form $g\mu_{\mathrm{B}}\mathbf{B}_{\mathrm{SO}}(\mathbf{k})\boldsymbol{\sigma}/2$. From eq. (9.6) we find

$$\mathbf{B}_{\mathrm{SO}}(\mathbf{k}) = \frac{2}{g\mu_{\mathrm{B}}}\begin{pmatrix} \alpha_{\mathrm{R}}k_y + \beta_{\mathrm{D}}k_x \\ -\alpha_{\mathrm{R}}k_x - \beta_{\mathrm{D}}k_y \\ 0 \end{pmatrix}. \tag{9.16}$$

This field is oriented in the plane of the electron gas with its specific direction given by the vector $\mathbf{k}$. While the Rashba field is oriented

	GaAs	AlAs	InAs	InSb	CdTe	ZnSe
α (eÅ^2)	5.206	−0.243	117.1	523.0	6.930	1.057
β (eVÅ^3)	27.58	18.53	27.18	760.1	43.88	14.29

Table 9.1 Values of Rashba and Dresselhaus coefficients for certain materials (Winkler, 2003).

normal to the direction of motion, the Dresselhaus field has a more complicated dependence of its orientation on $\mathbf{k}$. The direction of $\mathbf{B}_{\mathrm{SO}}$ defines the spin-precession axis if the spin is not in an eigenstate where it is aligned parallel or antiparallel to $\mathbf{B}_{\mathrm{SO}}$. If the direction of propagation is changed from $\mathbf{k}$ to $-\mathbf{k}$, $\mathbf{B}_{\mathrm{SO}}$ reverses its sign leading to precession in the opposite direction.

In order to obtain a feeling for the consequences of the two additional terms on the eigenstates and eigenenergies of the hamiltonian, we diagonalize eq. (9.6) neglecting the Dresselhaus term. For H_0 we use a diagonal matrix with the in-plane kinetic energy on the diagonal. We then find the dispersion

$$E_{\pm} = \frac{\hbar^2 k_{\parallel}^2}{2m^{\star}} \pm \alpha_{\mathrm{R}} k_{\parallel},$$

with $k_{\parallel} = \sqrt{k_x^2 + k_y^2}$. These two branches of the dispersion are still parabolae, but their minima are shifted in $k_{\parallel}$ compared to the spin-degenerate dispersion by $\pm m^{\star}\alpha_{\mathrm{R}}/\hbar^2$. Figure 9.13 shows the two dispersions schematically. The minimum of the dispersion lies along a circle in the k_x-k_y-plane. The surfaces of constant energy are circles with different radii reminiscent of a two-subband system. A corresponding calculation considering only the Dresselhaus term in the hamiltonian (9.6) gives the same dispersion if α_{R} is replaced by β_{D} [see Fig. 9.14(a)–(c)]. If Rashba and Dresselhaus terms are of similar importance, the dispersion relation is strongly modified, as shown in Fig. 9.14(d).

We can estimate the importance of the spin–orbit interaction by comparing the spin splitting of the two branches of the dispersion at the Fermi energy, $\Delta E_{\mathrm{SO}} = 2\alpha_{\mathrm{R/D}} k_{\mathrm{F}}$, with the Fermi energy $E_{\mathrm{F}} = \hbar^2 k_{\mathrm{F}}^2/2m^{\star}$.

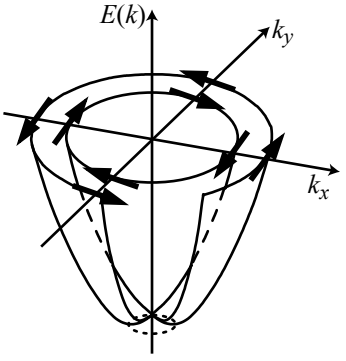


Fig. 9.13 Schematic representation of the in-plane dispersion taking the Rashba term into account. The arrows on circles of constant energy give the directions of the spin states (Winkler, 2003).

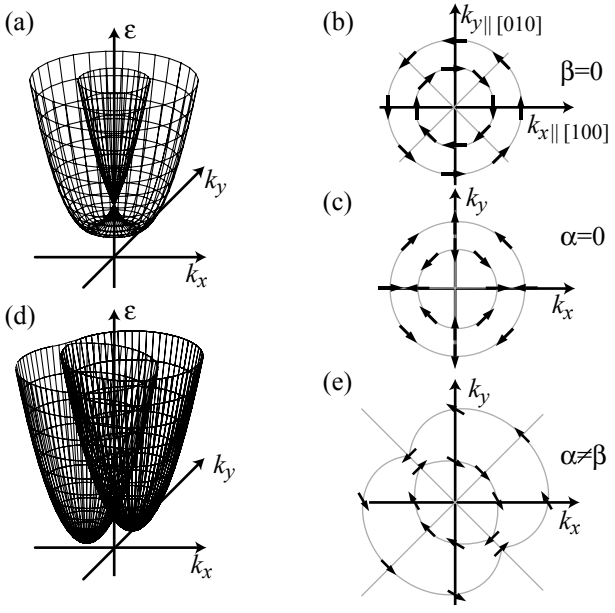


Fig. 9.14 Schematic presentation of the dispersion relations in the conduction band in the presence of spin–orbit interaction. (a) The case considering either only the Rashba or only the Dresselhaus term. (b) Fermi circles and spin orientations for the Rashba term. (c) Fermi contours and spin orientations for the Dresselhaus term. (d) Dispersion relation if the Rashba and Dresselhaus terms are of comparable magnitude. (e) Fermi contour and spin orientations for comparable Rashba and Dresselhaus terms. (Reprinted with permission from Ganichev *et al.*, 2004. Copyright 2004 by the American Physical Society.)

This gives

$$\frac{\Delta E_{\mathrm{SO}}}{E_{\mathrm{F}}} = \frac{4\alpha_{\mathrm{R/D}} m^\star}{\hbar^2 k_{\mathrm{F}}}.$$

For the effect to become important, apart from a big Rashba or Dresselhaus coefficient, a big mass and a small electron density (small k_{F}) is advantageous. As an example, we take the Dresselhaus effect in GaAs with an electron density $n_{\mathrm{s}} = 3 \times 10^{15}$ m^{-2} and a quantum well width of 100 Å. We find $\Delta E_{\mathrm{SO}}/E_{\mathrm{F}} \approx 0.07$. The Rashba effect in the same system is, for a (relatively strong) mean field $\langle E_z \rangle = 10^{-3}$ V/Å about five times smaller.

For each wave vector $\mathbf{k}$ a *local* quantization axis can be found in $\mathbf{k}$-space, along which the corresponding eigenstate is either $\uparrow$ or $\downarrow$. The direction of this axis varies with $\mathbf{k}$ such that, upon averaging all states, the same contributions arise from $\uparrow$ and $\downarrow$, and no net spin polarization arises. As a consequence, the magnetic moment vanishes. The spin orientations along circles of constant energy are shown as arrows in Fig. 9.13 for the Rashba splitting [see also Fig. 9.14(b)]. In the case of the Dresselhaus term the spin orientations look completely different, as shown in Fig. 9.14(c). The spin orientation in the presence of both contributions is depicted in Fig. 9.14(e) for spins at the Fermi energy.

Experimentally, attempts were made to demonstrate the tunability of the Rashba parameter α_{R} by applying gate voltages creating electric fields normal to the plane of the electron gas and measuring beating effects in the oscillatory magnetoresistance due to two different densities in the two spin subbands [see, e.g., Das *et al.* 1989; Luo *et al.* 1990; Nitta *et al.* 1997; Engels *et al.* 1997; Heida *et al.* 1998; Hu *et al.* 1999; Grundler 2000]. Other measurements (Brosig *et al.*, 1999), however, have not shown this beating. The importance of Rashba and Dresselhaus terms can also be measured via a coherent electron interference effect called weak antilocalization observable at weak magnetic fields normal to the electron gas (Knap *et al.*, 1996; Miller *et al.*, 2003). Measurements of the spin-galvanic photocurrent (Ganichev *et al.*, 2004) have permitted the measurement of the ratio $\alpha_{\mathrm{R}}/\beta_{\mathrm{D}} = 2.15$ for an InAs quantum well. Also Raman spectroscopy can provide quantitative information about spin–orbit interaction in two-dimensional systems (Jusserand *et al.*, 1995). The coefficients α_{R} and β_{D} have recently been measured by a variant of time-resolved Faraday rotation (Meier *et al.*, 2007).

9.7 Summary of characteristic quantities

In the following we summarize the important characteristic quantities of a two-dimensional electron gas in gated heterostructures.

Electron density. The sheet density n_{s} of electrons is determined by the applied gate voltage via the field effect. Typical values are $10^{11} - 10^{12}$ cm^{-2}.

Dispersion relation. The dispersion relation in a GaAs heterostructure is parabolic for the motion in the plane, i.e.,

$$E_{n\mathbf{k}} = E_n + \frac{\hbar^2 \mathbf{k}^2}{2m^\star}.$$

The wave vector $\mathbf{k} = (k_x, k_y)$ is in-plane.

Wave function. Each state $(n, \mathbf{k})$ comes with an (envelope) wave function

$$\psi_{n\mathbf{k}}(x, y, z) = \chi_n(z) e^{i(k_x x + k_y y)}.$$

The function χ_n describes the quantization in the z-direction, whereas the plane waves represent the free in-plane motion.

Density of states. From the parabolic dispersion, a constant density of states

$$\mathcal{D}_{2\mathrm{D}}(E) = \frac{g_\mathrm{s} g_\mathrm{v} m^\star}{2\pi\hbar^2},$$

follows for each subband n, where g_s describes the degree of spin degeneracy ($g_\mathrm{s} = 2$ in n-GaAs heterostructures and in Si-MOSFETs) and g_v is the number of degenerate conduction band minima ($g_\mathrm{v} = 1$ in GaAs heterostructures and $g_\mathrm{v} = 2$ in Si-MOSFETs).

Quantum limit. A two-dimensional electron gas with only one quantized subband occupied is said to be in the quantum limit. This is the closest physical realization of a mathematically ideal two-dimensional system. Often, the term 'two-dimensional electron gas' denotes a two-dimensional electron gas in the quantum limit. In the following, we will also use it in this sense.

Fermi energy and electron density. The electron density in a two-dimensional electron gas (in the quantum limit) is related to the Fermi energy via

$$n_\mathrm{s} = \mathcal{D}_{2\mathrm{D}} \cdot (\mu - E_0) = \mathcal{D}_{2D} \cdot E_\mathrm{F},$$

where μ is the electrochemical potential of the electron gas and E_0 is the quantization energy of the ground state subband. The difference $E_\mathrm{F} = \mu - E_0$ is called the Fermi energy, in analogy with metallic systems.

Fermi wave vector, Fermi energy and electron density. A Fermi wave vector k_F can be defined via the dispersion relation leading to the relations

$$k_\mathrm{F} = \sqrt{\frac{2m^\star E_\mathrm{F}}{\hbar^2}} = \sqrt{\frac{4\pi n_\mathrm{s}}{g_\mathrm{s} g_\mathrm{v}}}.$$

Fermi wavelength. The Fermi wavelength resulting from the general relation $k = 2\pi/\lambda$ is

$$\lambda_\mathrm{F} = \frac{2\pi}{k_\mathrm{F}} = \sqrt{\frac{g_\mathrm{s} g_\mathrm{v} \pi}{n_\mathrm{s}}}.$$

It is of the order of the mean electron separation $1/\sqrt{n_\mathrm{s}}$.

Fermi velocity. The (group) velocity of an electron at the Fermi energy is given by

$$v_{\mathrm{F}} = \frac{\hbar k_{\mathrm{F}}}{m^\star}.$$

Bohr radius. The Bohr radius is the characteristic length scale of the Coulomb interaction in the electron gas. It is a material-specific quantity given by

$$a_{\mathrm{B}}^\star = \frac{4\pi\epsilon\epsilon_0\hbar^2}{m^\star e^2}.$$

Rydberg energy. The Rydberg energy is the characteristic energy scale of the Coulomb interaction in the electron gas given by

$$E_{\mathrm{Ry}}^\star = \frac{e^4 m^\star}{2(4\pi\epsilon\epsilon_0)^2\hbar^2}.$$

Thomas–Fermi screening length and Thomas–Fermi wave vector. The Thomas–Fermi screening length and the corresponding wave vector are, in two-dimensional electron gases, given by

$$\lambda_{\mathrm{TF}} = \pi a_{\mathrm{B}}^\star \quad q_{\mathrm{TF}} = \frac{2}{a_{\mathrm{B}}^\star}.$$

independent of the electron density.

Nonlinear screening length. The nonlinear screening length describes the characteristic length scale for the separation of the homogeneous electron gas into separate electron puddles upon reducing the density. It is given by

$$R_{\mathrm{c}} = \frac{\sqrt{N_{\mathrm{d}}}}{n_{\mathrm{s}}}.$$

Percolation threshold. Below a critical density n_{c} in the electron gas there is no connected domain extending over the whole macroscopic sample. This density is given by

$$n_{\mathrm{c}} = 0.11\frac{\sqrt{N_{\mathrm{d}}}}{s},$$

where s is the separation between the doping plane (here assumed to be a δ-doping layer) and the electron gas.

Further reading

- Two-dimensional electron gases: Davies 1998; Ando *et al.* 1982.
- Spin–orbit interaction: Winkler 2003.
- Papers: Ando 1982; Stern and Das Sarma 1984.

Exercises

(9.1) Consider the capacitance between the two-dimensional electron gas in a heterostructure and a large area metallic top gate as given by eq. (9.1). You intend to measure the density of states of the electron gas as a function of the Fermi energy. How could you optimize the geometrical design of the structure for this purpose, and what would your measurement setup look like?

(9.2) Suppose you intend to measure the density of states of graphene as a function of Fermi energy by measuring the capacitance between a single-layer graphene sheet and a highly doped silicon substrate with silicon dioxide of thickness d between silicon and the sheet. Make a suitable sketch of the electrostatic problem. Write down your electrostatic considerations and demonstrate how the density of states of graphene can be extracted.

(9.3) Consider a two-dimensional electron gas (2DEG) sandwiched between a highly doped back gate in the substrate, and a metallic top gate. The separation between the electron gas and the back gate is d_{BG}, and that between the electron gas and the top gate is d_{TG}. Draw a sketch of this arrangement. You ground the two-dimensional electron gas and apply a small low frequency modulated voltage between the electron gas and the back gate. If the electron gas were a perfect metal, no modulated electric field would reach the top gate. Work out how this situation is different in a real two-dimensional electron gas with its finite density of states. How can this setup be used for measuring the density of states in the two-dimensional electron gas? Discuss advantages and disadvantages of this method compared to the measurement of the density of states via the gate–2DEG capacitance.

(9.4) In this problem, you deepen your understanding of Friedel oscillations. To this end, consider an electron gas in the one-dimensional potential

$$V(x) = \begin{cases} 0 & \text{for } x > 0 \\ \infty & \text{for } x \leq 0 \end{cases} .$$

(a) What happens if a particle with energy E hits this potential barrier from the right? Calculate the transmission and reflection coefficients and probabilities.

(b) What are the wave functions $\psi_k(x)$ for $x > 0$?

(c) Normalize the wave function on a large one-dimensional volume L so that

$$\int_0^L dx\, |\psi_k(x)|^2 = 1 + \mathcal{O}\left(\frac{1}{kL}\right).$$

(d) At zero temperature all the states up to k_{F} are occupied. Calculate the local one-dimensional electron density at zero temperature.

(e) Sketch the resulting electron density. How does it depend on k_{F} at large distances from the barrier?

This page intentionally left blank

Diffusive classical transport in two-dimensional electron gases

10

10.1 Ohm's law and current density

Figure 10.1 shows schematically a setup for the measurement of electron transport in a sample. A current source drives the current I through the sample. A voltage builds up along the direction of the current flow, which can be measured using a voltmeter. The basic law of electron transport is Ohm's law, found in 1826 by Georg Simon Ohm. It states that the current I driven through the sample, and the longitudinal voltage drop U are proportional, i.e.,

$$U = RI. \tag{10.1}$$

The proportionality constant R is called electrical resistance and has dimensions $1\,\Omega = 1\,\mathrm{V/A}$. Its inverse, $G = R^{-1}$, is called electrical conductance and has the dimensions $1\,\mathrm{S} = 1\,\mathrm{A/V}$.

Ohm's law was extended in 1845 by Gustav Kirchhoff to DC (direct current) networks of ohmic resistors. Today, we know his two laws as Kirchhoff's current law and Kirchhoff's voltage law. The former states that the sum of all electrical currents I_n flowing into a junction is zero:

$$\sum_n I_n = 0.$$

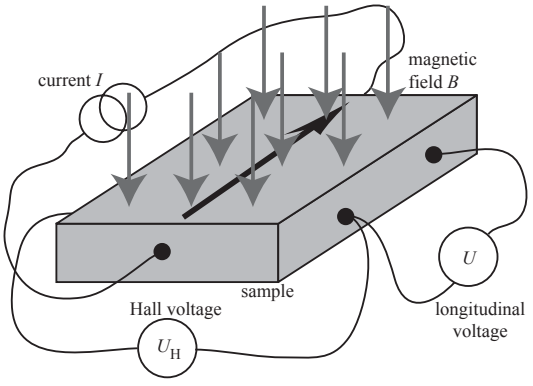


Fig. 10.1 Schematic view of a resistance measurement. A current source drives the current I through the sample. The voltage drop U is measured along the edge of the sample in the direction of current flow. If a magnetic field B is applied normal to the current flow, a Hall voltage U_H builds up normal to the current flow and normal to the magnetic field direction.

In its expression, currents flowing into the junction are counted negative, those flowing away from the junction positive. Nowadays we know that this law is a direct consequence of the conservation of electric charge as it is descibed by the continuity equation within Maxwell's theory of electromagnetism. Kirchhoff's voltage law states that the sum of all voltage drops U_m across components within a circuit loop add to zero:

$$\sum_m U_m = 0$$

Within Maxwell's theory, this loop rule can be derived from Faraday's law of induction.

In the following, we will go beyond these macroscopic laws of electric circuits and discuss the physical background and the microscopic origin of the electrical resistance. Because samples made of the same material, but with different geometries (e.g., wires of different lengths), will have different resistances, this quantity is not an appropriate quantity for the description of the transport properties of the material. For a homogeneous isotropic material, Ohm's law is therefore frequently written in the (local) form

$$\mathbf{j}(\mathbf{r}) = \sigma \mathbf{E}(\mathbf{r}), \tag{10.2}$$

where $\mathbf{j}(\mathbf{r})$ is the electrical current density, $\mathbf{E}(\mathbf{r})$ is the electric field, and σ is the electrical conductivity. In a homogeneous conductor, the latter does not depend on $\mathbf{r}$.

Three-dimensional systems. In a three-dimensional system, σ can be a 3×3 tensor; in the case of an isotropic homogeneous conductor it reduces to a scalar quantity. The current density $\mathbf{j}$ denotes the number of charges traversing a unit area normal to the current flow within a time unit. If the sample is a cuboid of length L, width W, and thickness d, then the current is $I = jWd$, and the electric field is related to the voltage drop via $E = U/L$. Therefore, the conductance is related to conductivity by $G = \sigma Wd/L$. The specific resistivity $\rho = \sigma^{-1}$ is the inverse of the conductivity. It is related to the resistance via $R = \rho L/Wd$. In general, the specific resistivity and the conductivity of a material can be determined if the geometry of the sample is known.

Two-dimensional systems. In homogeneous two-dimensional systems, such as heterostructures with a two-dimensional electron gas, Ohm's phenomenological law remains valid. The current density in two-dimensional systems is, however, defined as $j = I/W$, where W is the width of the sample. In two dimensions, the current density therefore has the units $[j] = \mathrm{A/m}$ (instead of $\mathrm{A/m^2}$ in three dimensions). Correspondingly, the units of the conductivity are $[\sigma] = \Omega^{-1}$ (rather than $\Omega^{-1}\mathrm{m}^{-1}$ in three dimensions), and those of the specific resistivity are $[\rho] = \Omega$ (rather than $\Omega\mathrm{m}$). In anisotropic two-dimensional conductors, or in the presence of a magnetic field, the conductance and the specific resistivity are 2×2 tensors.

Diffusive transport regime. Different materials can have the same specific resistance. In the following, we will see that the specific resistance can be expressed as a function of the density of charge carriers n participating in electron transport, and a scattering time τ of the charge carriers. In the *diffusive transport regime* the scattering of electrons takes place on length scales that are small compared to the size of the sample.

10.2 Hall effect

With the discovery of the Hall effect in 1879 (Hall, 1879; Hall, 1880), Edwin Herbert Hall founded the field of magnetotransport phenomena (also called galvanomagnetic effects) comprising all effects that an external homogeneous magnetic field causes in a conducting sample through which an electric current is driven. Hall found out that a magnetic field normal to the direction of current flow results in a voltage U_H between two points with their connecting line normal to the magnetic field and normal to the current flow (Fig. 10.1). Hall found this effect about 20 years prior to the discovery of the electron by Sir Joseph John Thomson in 1897. This phenomenon is called Hall effect. The Hall voltage U_H is proportional to the applied magnetic field B and to the current I through the sample, i.e.,

$$U_\mathrm{H} = R_\mathrm{H} B I. \tag{10.3}$$

The constant of proportionality R_H is called the Hall coefficient. Figure 10.2 shows the five data points measured by Hall on a thin gold foil. The horizontal axis is the magnetic field scale, the vertical axis is proportional to the Hall resistance U_H/I. Hall found out that the longitudinal voltage U is independent of the magnetic field strength, in contrast to the Hall voltage.

Drude model. As with Ohm's law, the Hall effect can be also described within the Drude model of electrical conduction which will be discussed

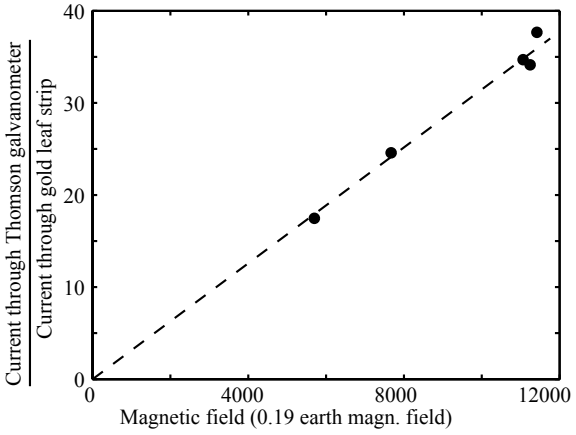


Fig. 10.2 Original data of E.H. Hall compiled from his 1879 papers (Hall, 1879; Hall, 1880).

in detail in the next section. Within this model, the Hall coefficient R_{H} of a three-dimensional sample is related to the electron density n:

$$R_{\mathrm{H}}^{\mathrm{3D}} = -\frac{1}{n|e|d}. \tag{10.4}$$

Here, d is the sample thickness in the magnetic field direction. The minus sign originates from the negative charge of the electrons. Thin conducting films are therefore well suited for Hall measurements. In metals, the Hall voltage is typically very small, because the electron density is very large. In contrast, large Hall voltages can be reached in semiconductors due to the small achievable electron densities. In two-dimensional electron gases, the density of carriers is a sheet concentration n_{s} with dimensions m^{-2} and therefore the thickness d in the above expression is irrelevant. Therefore, in two-dimensional systems the Hall coefficient is

$$R_{\mathrm{H}}^{\mathrm{2D}} = -\frac{1}{n_{\mathrm{s}}|e|}. \tag{10.5}$$

The Hall effect can be incorporated in the tensor representation of the electrical conductivity σ in eq. (10.2). If the magnetic field is applied in the z-direction, the Hall effect appears in the matrix elements σ_{xy} and σ_{yx}.

10.3 Drude model with magnetic field

From a microscopic point of view, scattering processes give rise to the appearance of electrical resistance. Possible scattering mechanisms are, for example, scattering at crystal defects, charged impurities, or lattice vibrations (phonons). As an example, we consider electrons near the Γ-point of the GaAs conduction band having the (effective) mass $m^\star$ and charge $-|e|$. We characterize the scattering processes of the electrons with a mean scattering time τ.

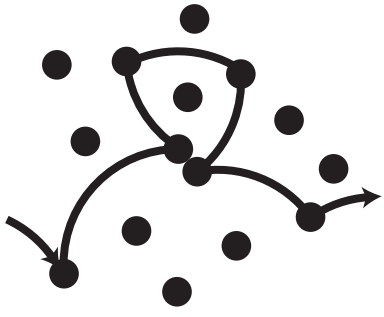


Fig. 10.3 Example for a diffusive electron trajectory in a magnetic field.

Figure 10.3 shows the diffusive trajectory of an electron if a magnetic field is applied normal to the plane of electron motion. Between scattering events, the electron moves on circular orbits. After a collision, the electron has a direction of motion completely uncorrelated with that before the collision. We describe the situation in the following way. We assume that an electric field $\mathbf{E} = (E_x, 0, 0)$ is applied in the x-direction and that the magnetic field $\mathbf{B} = (0, 0, B_z)$ is in the z-direction, normal to the average current flow (see Fig. 10.1). We consider the electronic motion to be quantized in the z-direction due to the confinement in the heterostructure, but describe the motion in the plane semiclassically. In the plane of the two-dimensional electron gas, the Lorentz force $\mathbf{F} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B})$ acts on the electrons and Newton's equations describing the electron motion between two collisions are

$$m^\star \frac{dv_x}{dt} = -|e|(E_x + v_y B_z) \quad \text{and} \quad m^\star \frac{dv_y}{dt} = +|e| v_x B_z.$$

Introducing the cyclotron frequency $\omega_c = |e|B_z/m^\star$ and the drift velocity $v_D = E_x/B_z$ these equations can be written as

$$\frac{dv_x}{dt} = -\omega_c(v_D + v_y) \quad \text{and} \quad \frac{dv_y}{dt} = +\omega_c v_x. \tag{10.6}$$

The solution of the equations of motion is

$$\mathbf{v}(t) = \begin{pmatrix} 0 \\ -v_D \end{pmatrix} - \begin{pmatrix} v_y(0) + v_D \\ -v_x(0) \end{pmatrix} \sin\omega_c t + \begin{pmatrix} v_x(0) \\ v_y(0) + v_D \end{pmatrix} \cos\omega_c t.$$

We have assumed that the last scattering event took place at time $t = 0$ leaving the electron with the velocity $(v_x(0), v_y(0))$. The vectors in front of the sin and the cos terms are orthogonal to each other. As a consequence, the last two terms describe a circular motion with frequency ω_c, the so-called cyclotron motion (motion on a cyclotron orbit). The first term describes the drift of the orbit center in the y-direction. This means that the time-averaged electron motion is in the negative y-direction normal to the direction of the magnetic and electric fields, despite the electric field pushing it in the x-direction. This motion is therefore called $\mathbf{E} \times \mathbf{B}$-drift.

The cyclotron frequency introduces a time scale into the problem which will compete with the mean scattering time τ. For $\omega_c\tau \ll 1$, i.e. for small magnetic fields, the electrons cannot complete a cyclotron orbit without being scattered. In the opposite case $\omega_c\tau \gg 1$, i.e. for large magnetic fields, this is well possible.

Over large time scales $t \gg \tau$ the electron will have a mean drift velocity in the plane. In order to calculate this drift velocity, we assume that after each collision, the electron has a velocity of magnitude v_0 in a random direction which is uncorrelated to the direction before the collision. We express this mathematically by introducing the probability $\mathcal{P}(\varphi)d\varphi$ that the electron moves in the direction φ immediately after a collision. Our assumption means that the probability distribution for the direction of the initial velocity $\mathbf{v}(0)$ is given by

$$\mathcal{P}(\varphi)d\varphi = \frac{d\varphi}{2\pi}. \tag{10.7}$$

In order to calculate the drift velocity, we further need the probability $w\,dt$ that an electron is scattered within an infinitesimal time interval dt. If $P(t)$ is the probability that the electron has not been scattered until time t after the last collision, we have the rate equation $dP(t)/dt = -wP(t)$ with the solution $P(t) = Ce^{-wt}$, where the constant C can be derived from the initial condition $P(0) = 1$ to be $C = 1$. The product $p(t) = P(t)w\,dt$ is the probability that the electron scatters within the time interval $[t, t + dt]$ after the last collision. Therefore, the mean scattering time is

$$\tau = \int_0^\infty dt\, we^{-wt}t = \frac{1}{w},$$

and

$$p(t)dt = e^{-t/\tau}\frac{dt}{\tau}. \tag{10.8}$$

The mean drift velocity is now obtained from eq. (10.6) by multiplying the velocity $\mathbf{v}(t)$ with the probability $\mathcal{P}(\varphi)d\varphi$ for starting at angle φ and with $p(t)\,dt$ for undisturbed motion during t and integrating over all times $t \geq 0$ and angles φ. Mathematically this means

$$\bar{v} = \int_0^{2\pi} d\varphi\, \mathcal{P}(\varphi) \int_0^{\infty} dt\, p(t)\mathbf{v}(t). \tag{10.9}$$

This procedure gives, for the components of the drift velocity,

$$\begin{aligned} \bar{v}_x &= -v_\mathrm{D}\frac{\omega_\mathrm{c}\tau}{1+\omega_\mathrm{c}^2\tau^2} = -E_x\frac{|e|\tau/m^\star}{1+\omega_\mathrm{c}^2\tau^2} \\ \bar{v}_y &= -v_\mathrm{D}\frac{\omega_\mathrm{c}^2\tau^2}{1+\omega_\mathrm{c}^2\tau^2} = -v_\mathrm{D}\left(1-\frac{1}{1+\omega_\mathrm{c}^2\tau^2}\right) \\ &= -v_\mathrm{D} + \Delta\bar{v}_y. \end{aligned} \tag{10.10}$$

Performing the averaging, we have incorporated the effect of scattering into the description. Obviously, scattering reduces the drift of the cyclotron orbit center in the y-direction by $\Delta\bar{v}_y = v_\mathrm{D}/(1+\omega_\mathrm{c}^2\tau^2) = \bar{v}_x/\omega_\mathrm{c}\tau$. This frictional contribution can be understood in the following way: drift motion with the velocity $-v_\mathrm{D}$ leads to a frictional force $F_y = m^\star v_\mathrm{D}/\tau$. Its effect corresponds to an effective electric field $E_y = -m^\star v_\mathrm{D}/|e|\tau$. According to the first equation this has to lead to a change in velocity $\Delta\bar{v}_y = -E_y|e|\tau/m^\star(1+\omega_\mathrm{c}^2\tau^2)$ which is exactly $\Delta\bar{v}_y = v_\mathrm{D}/(1+\omega_\mathrm{c}^2\tau^2)$. For $B_z \to 0$, $\bar{v}_y \to 0$ and $\bar{v}_x$ goes to a finite zero field limit governed by the action of the applied electric field. The proportionality constant between the drift velocity $\bar{v}_x$ at zero magnetic field and the electric field

$$\mu = \frac{|e|\tau}{m^\star} \tag{10.11}$$

is called the electron mobility.

We see from eq. (10.10) that the mean drift velocity at finite magnetic field is not parallel to the direction of the electric field (x-direction), but encloses with the x-direction the so-called Hall angle θ obeying

$$\tan\theta = \frac{\bar{v}_y}{\bar{v}_x} = \omega_\mathrm{c}\tau = \mu B.$$

For $B_z \to 0$, the Hall angle goes to zero and the electron drift and electric field are oriented in the same direction.

The current density can now be calculated from the drift velocity to be

$$\begin{aligned} j_x &= -n_\mathrm{s}|e|\bar{v}_x = \frac{n_\mathrm{s}e^2\tau}{m^\star}\frac{1}{1+\omega_\mathrm{c}^2\tau^2}E_x \\ j_y &= -n_\mathrm{s}|e|\bar{v}_y = \frac{n_\mathrm{s}e^2\tau}{m^\star}\frac{\omega_\mathrm{c}\tau}{1+\omega_\mathrm{c}^2\tau^2}E_x. \end{aligned}$$

Figure 10.4 schematically shows the situation in a so-called Hall bar geometry. The geometry allows current flow only in the x-direction.

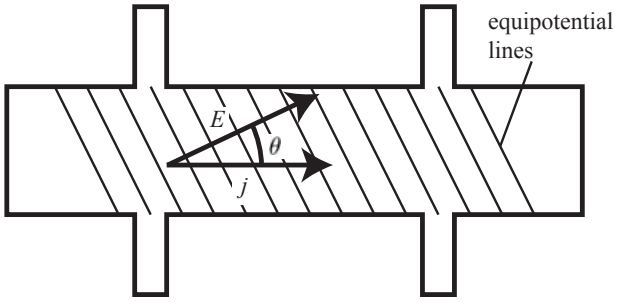


Fig. 10.4 Relative direction of the current density vector and the electric field at finite magnetic field. The two vectors span the Hall angle θ. The equipotential lines normal to the direction of **E** show how the classical Hall voltage comes about.

The vectors **j** and **E** enclose the Hall angle θ. The equipotential lines are not normal to the direction of the current flow. As a result, a finite Hall voltage is measured between corresponding contacts on opposite sides of the bar.

At finite magnetic field normal to the plane of the electron gas, the conductivity is a 2×2 tensor. We define the components

$$\sigma_{xx}(B) = \frac{n_s e^2 \tau}{m^\star} \frac{1}{1+\omega_c^2\tau^2} \tag{10.12}$$

$$\sigma_{xy}(B) = \frac{n_s e^2 \tau}{m^\star} \frac{\omega_c\tau}{1+\omega_c^2\tau^2} = \omega_c\tau\sigma_{xx} = \frac{n_s|e|}{B} - \frac{\sigma_{xx}}{\omega_c\tau}. \tag{10.13}$$

The same calculation performed for an electric field in the y-direction gives tensor components $\sigma_{yy} = \sigma_{xx}$ and $\sigma_{yx} = -\sigma_{xy}$ and the whole conductivity tensor is determined. The components σ_{xx} and σ_{xy} are plotted as a function of magnetic field in Fig. 10.5.

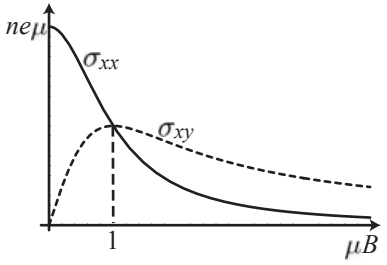


Fig. 10.5 Drude conductance as a function of magnetic field, expressed as $\mu B = \omega_c\tau$. At $\mu B = 1$, $\sigma_{xx} = \sigma_{xy}$.

The relation between the current density and the electric field can now be written as

$$\begin{pmatrix} j_x \\ j_y \end{pmatrix} = \begin{pmatrix} \sigma_{xx} & -\sigma_{xy} \\ \sigma_{xy} & \sigma_{xx} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix}. \tag{10.14}$$

The tensor of the specific resistivity is obtained by tensor inversion resulting in

$$\rho_{xx} = \frac{\sigma_{xx}}{\sigma_{xx}^2+\sigma_{xy}^2} = \frac{m^\star}{n_s e^2\tau} \tag{10.15}$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_{xx}^2+\sigma_{xy}^2} = \frac{B}{|e|n_s}. \tag{10.16}$$

The magnetic field dependence of these two components is plotted in Fig. 10.6. The resistivity tensor relates the electric field and the current density via

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} \rho_{xx} & \rho_{xy} \\ -\rho_{xy} & \rho_{xx} \end{pmatrix} \begin{pmatrix} j_x \\ j_y \end{pmatrix}. \tag{10.17}$$

As originally observed by Hall, the longitudinal component ρ_{xx} of the specific resistivity is independent of magnetic field. It is in fact determined by the scattering time τ. The transverse component ρ_{xy}, which is also called the Hall resistivity, is independent of τ, but linear in B.

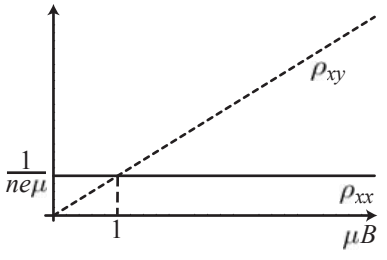


Fig. 10.6 Specific resistivity in the Drude model as a function of magnetic field. For $\mu B = 1$, $\rho_{xx} = \rho_{xy}$.

A measurement of the two independent components of the resistivity tensor allows us to determine the density and the mobility (scattering time) of the electron gas. The electron density is obtained from the measurement of the Hall resistivity via

$$n_{\mathrm{s}} = \frac{1}{|e|\, d\rho_{xy}/dB|_{B=0}} \tag{10.18}$$

and the electron mobility is given by

$$\mu = \frac{d\rho_{xy}/dB|_{B=0}}{\rho_{xx}(B=0)}. \tag{10.19}$$

The scattering time is then determined from eq. (10.11).

The Hall angle has been made visible in an experiment (Novak *et al.*, 1998). Thin layers of very weakly volume doped n-GaAs with charge carrier concentrations of about $1.3 \times 10^{15}\,\mathrm{cm}^{-3}$ were prepared. Rectangular samples with two 6 mm-wide ohmic contacts with a separation of 1.65 mm were fabricated. Measurements were performed at a temperature $T = 4.2\,\mathrm{K}$. The linear conductance of the samples is very small. At sufficiently high applied voltages (fields of a few V/cm), an electrical breakthrough occurs in which neutral donors are ionized by electronic collisions. Their electrons are then available for electron transport. This breakthrough occurs within current filaments which have a constant conductivity in their center described by the Drude model. The current carrying filaments can be made visible by spatially resolved measurement of the photoluminescence of the samples. To this end the experimentalists illuminated the sample with red LEDs and observed the excitonic recombination of the electrons and holes with a suitable camera. The intensity of the excitonic recombination is suppressed by the presence of free charge carriers in the conducting filaments leading to dark stripes as shown in Fig. 10.7. If a magnetic field is applied normal to the direction of the current, the orientation of the filaments changes by the Hall angle θ.

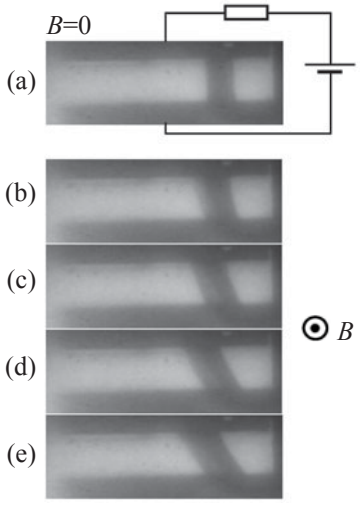


Fig. 10.7 Spatially resolved photoluminescence images of an n-GaAs sample with electric breakthrough. (a) A dark stripe between the ohmic contacts indicates a current filament in which conductance is described by the Drude model. (b)–(e) With increasing magnetic field the direction of the filament changes. The angle between the original direction at $B = 0$ and the direction at finite B is the Hall angle. The applied magnetic fields are $B = 64\,\mathrm{mT}$ (b), $B = 126\,\mathrm{mT}$ (c), $B = 188\,\mathrm{mT}$ (d), $B = 251\,\mathrm{mT}$ (e). The current through the sample is 1 mA. (Reprinted with permission from Novak *et al.*, 1998. Copyright 1998 by the American Physical Society.)

Crucial for this experiment is that the direction of the electric field is determined by the boundary conditions at the contacts, while the direction of the current can follow the Hall angle. The geometry of the sample is therefore an important ingredient. In the following section we will discuss the influence of sample geometries on the measured resistance.

10.4 Sample geometries

We have seen in the previous section how the conductivity and the specific resistivity of a two-dimensional electron gas are related to the material-specific quantities n and τ. Now we will discuss how the specific resistivity tensor can actually be measured.

The geometry of the sample and the measurement setup may have a crucial influence on the measured electrical resistance. Suitable sample

geometries have therefore been developed that allow us to determine the specific resistivity from the measured resistance reliably and easily. In principle, the arrangement of contacts and sample edges determines the current density and electric field distribution in a sample. In the case of a DC (direct current, i.e., zero or low-frequency) measurement, these two vector fields are determined from the two Maxwell equations

$$\nabla \mathbf{E} = 0 \tag{10.20}$$

$$\nabla \times \mathbf{E} = 0 \tag{10.21}$$

and Ohm's law, eq. (10.2). The vector field $\mathbf{E}$ is given by the two eqs. (10.20) and (10.21), and by the boundary condition that field components parallel to the edges of ohmic contacts vanish, i.e.,

$$E_{\parallel}(\mathbf{r}) = 0 \text{ for } \mathbf{r} \text{ on the boundary of an ohmic contact.}$$

At dielectric edges of a sample there are no boundary conditions for $\mathbf{E}$.

In order to find the equations for the vector field of the current density, we use the continuity equation $\partial\rho/\partial t + \nabla \mathbf{j} = 0$, from which we obtain in the stationary limit ($\partial\rho/\partial t = 0$)

$$\nabla \mathbf{j} = 0.$$

If we take the curl of both sides of Ohm's law and use the Maxwell equations for the electric field we obtain, for a homogeneous system with σ being constant in space,

$$\nabla \times \mathbf{j} = \nabla \times (\sigma \mathbf{E}) = \sigma \nabla \times \mathbf{E} = 0.$$

As with the electric field, the current density has zero divergence and zero curl. Complementing the boundary conditions for the electric field, we require at sample edges the current density component normal to the boundary to vanish, i.e.,

$$j_{\perp}(\mathbf{r}) = 0 \text{ for } \mathbf{r} \text{ on the dielectric boundary of the structure.}$$

The conductivity σ does not enter the equations for $\mathbf{j}$ and plays a role only via the boundary conditions. For example, the current density at the boundary to an ohmic contact is given by

$$j_{\perp}(\mathbf{r}) = \sigma_{xx} E_{\perp} \text{ for } \mathbf{r} \text{ on the edge of an ohmic contact,}$$

$$j_{\parallel}(\mathbf{r}) = \sigma_{xy} E_{\perp} \text{ for } \mathbf{r} \text{ on the edge of an ohmic contact.}$$

In the following, we will illustrate the consequences of the equations determining $\mathbf{j}$ and $\mathbf{E}$ with a few important examples.

Hall bar geometry. If a sample containing a two-dimensional electron gas is made to have the shape of a long bar with an ohmic contact at each end [Fig. 10.8(a)], the field lines for $\mathbf{j}$ and $\mathbf{E}$ are easily found. Both vector fields are homogeneous and directed parallel to the axis of the bar.

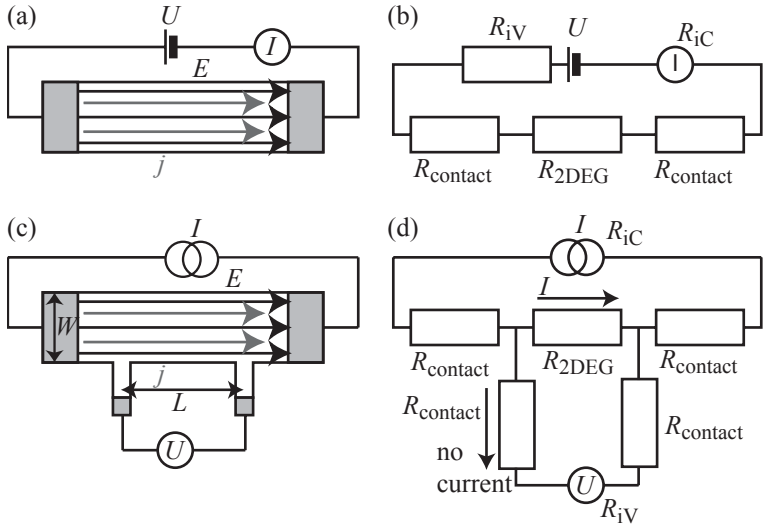


Fig. 10.8 (a) Two-terminal measurement with a bar geometry. (b) Equivalent circuit with contact resistances R_{contact} and the internal resistances of the voltage source R_{iV} (typically $< 50\,\Omega$) and of the ammeter R_{iC} (typically $\leq 10\,\Omega$). (c) Four-terminal measurement with a bar geometry. (d) Equivalent circuit with contact resistances R_{contact}, the internal resistance of the voltmeter R_{iV} (typically $> 10\,\mathrm{M}\Omega$) and the internal resistance of the current source (typically $> 10\,\mathrm{M}\Omega$).

In general, the setup shown schematically, which is called a *two-terminal measurement*, will not lead to the measurement of the two-dimensional electron gas resistance, because the resistances of the electrical contacts, R_{contact}, the internal resistance of the voltage source, R_{iV}, and that of the ammeter, R_{iC}, are connected in series. According to the equivalent circuit shown in Fig. 10.8(b) we obtain for the measured resistance $R = U/I = R_{2\mathrm{DEG}} + 2R_{\mathrm{contact}} + R_{\mathrm{iV}} + R_{\mathrm{iC}}$. A significant improvement can be achieved using the four-terminal arrangement depicted schematically in Fig. 10.8(c). Two narrow side contacts have been attached to the bar which leave the current distribution and the electric field essentially undisturbed, but allow the voltage to be picked up along the electron gas. In the measurement setup the voltage source has been replaced by a current source which delivers a well-defined current I independent of the size of the load resistance. The resistances of the voltage contacts do not play a role, because no current will flow through the voltmeter due to its very large internal resistance. As a consequence, the electric field in the two-dimensional electron gas is given by $|\mathbf{E}| = U/L$, the current density is $|\mathbf{j}| = I/W$ and therefore the specific resistivity takes the value

$$\rho_{xx} = \frac{U}{I}\frac{W}{L}.$$

If a magnetic field is applied normal to the plane of the two-dimensional electron gas, the pattern of field lines will be changed as depicted in Fig. 10.9. Near the current contacts which are equipotential lines of the electric field, the equipotentials are forced to run parallel to the edge of the contact, roughly as long as the distance from the contact is less than the width W of the sample. The field lines of the current density, however, must be at the Hall angle relative to the direction of the electric field. Far away from the current contacts (much further than W) the edges of the sample force the field lines of $\mathbf{j}$ into a direction parallel to

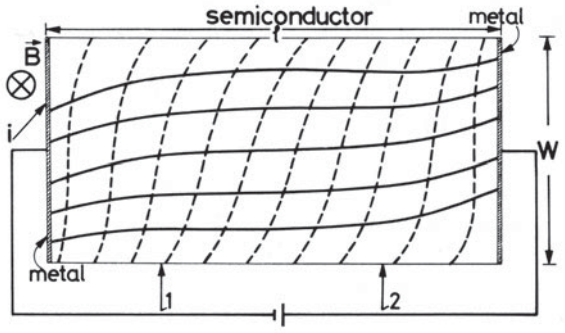


Fig. 10.9 Equipotential lines for the electric potential (dashed) and field-lines for the current density (solid) in a short bar shaped sample with two metal contacts (Seeger, 2004).

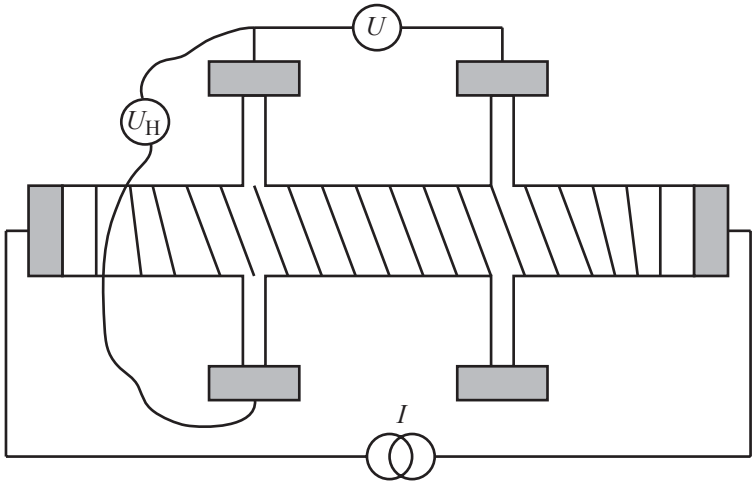


Fig. 10.10 Hall bar geometry with equipotential lines for the electric potential and field lines for the current density.

the sample edge. The field lines of **E**, however, must be at the Hall angle to the current density and therefore to the axis of the bar. This region will only form if the length of the bar between the current contacts is much larger than W (Fig. 10.10). As a rule of thumb, the separation of a voltage probe and a current contact has to be more than $4W$. In regions that are a distance much larger than W from the current contacts, the proper Hall voltage can be measured normal to the axis of the bar. This is achieved in the *Hall bar* depicted in Fig. 10.10 by fabricating pairs of contacts at opposite sides of the bar. For this case one obtains

$$\rho_{xy} = \frac{U_\mathrm{H}}{I}.$$

Figure 10.11 shows the photograph of a structure used today for the characterization of two-dimensional electron gases. The large areas of the contacts minimize the contact resistance at low temperatures. Such structures are well known from transport measurements on three-dimensional doped semiconductors.

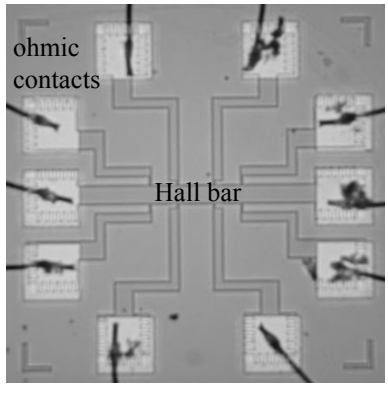


Fig. 10.11 Hall bar sample used today for characterizing two-dimensional electron gases. The Hall bar is 100 µm wide and about 1 mm long.

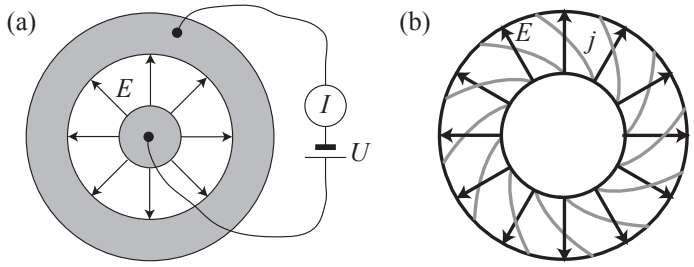


Fig. 10.12 (a) Corbino geometry for measuring the conductivity. (b) Electric field lines and current distribution in the Corbino geometry.

Corbino geometry. We have seen that a long and narrow Hall bar structure is crucial for the correct measurement of the resistivity tensor. Short and wide bars lead to a significant geometrical influence on the field line pattern (Fig. 10.9). This is very pronounced in the so-called Corbino geometry depicted in Fig. 10.12 (a). The cylindrical symmetry of this arrangement requires the electric field lines to point radially outwards. The field lines of the current density are logarithmic spirals of the form $\rho = ae^{\varphi/(\omega_c\tau)}$. Figure 10.12 (b) shows the field lines of $\mathbf{E}$ and $\mathbf{j}$ in the Corbino geometry. Only the current component flowing in radial direction is measured, i.e. the component σ_{xx} of the conductivity tensor. The radial current density at the inner contact is $j_\rho = I/(2\pi r_i)$, and the electric field strength is $E_\rho = j_\rho/\sigma_{xx} = I/(2\pi\sigma_{xx}r_i)$ there. The voltage difference between the inner and the outer contact is given by $U = I/(2\pi\sigma_{xx}) \ln r_a/r_i$. Neglecting possible contact resistances we find

$$\sigma_{xx} = \frac{I}{U}\frac{1}{2\pi}\ln\frac{r_a}{r_i}. \tag{10.22}$$

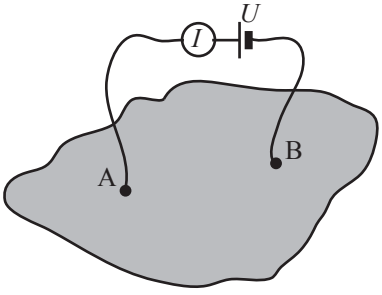


Fig. 10.13 Measurement between two points in an infinitely extended plane.

Resistance between two points in the plane. We now consider a two-dimensional electron gas in a plane (extended to infinity) into which we inject the current I_{AB} at point $\mathbf{r}_{\mathrm{A}}$ which will flow to infinitely remote contacts. The current density will be directed radially away from the contact due to the cylindrical symmetry of the problem and decay in the radial direction according to

$$j_\rho(\rho) = \frac{I_{\mathrm{AB}}}{2\pi\rho}.$$

At zero magnetic field, the electric field will also be directed radially and decay according to

$$E_\rho(\rho) = \frac{I_{\mathrm{AB}}}{2\pi\sigma_{xx}\rho}.$$

The vector fields $\mathbf{E}$ and $\mathbf{j}$ do not change if the point-like current injector is replaced by an injecting circular disk.

We now extend the arrangement by extracting the current at another point $\mathbf{r}_{\mathrm{B}}$ as shown in Fig. 10.13. As a result, we have a point-like current source and a point-like current sink in the plane which can be incorporated into the equations for the current density as

$$\begin{aligned} \nabla\mathbf{j} &= I_{\mathrm{AB}}[\delta(\mathbf{r}-\mathbf{r}_{\mathrm{A}}) - \delta(\mathbf{r}-\mathbf{r}_{\mathrm{B}})] \\ \nabla\times\mathbf{j} &= 0. \end{aligned}$$

The solution of these two equations can be found by introducing a potential function Λ defined via $\mathbf{j} = -\nabla\Lambda$. The potential function is determined by the two-dimensional Poisson equation

$$-\Delta\Lambda = I_{\mathrm{AB}}[\delta(\mathbf{r}-\mathbf{r}_{\mathrm{A}}) - \delta(\mathbf{r}-\mathbf{r}_{\mathrm{B}})]$$

with the solution

$$\Lambda = -\frac{I_{\mathrm{AB}}}{2\pi}\ln\left(\frac{|\mathbf{r}-\mathbf{r}_{\mathrm{A}}|}{|\mathbf{r}-\mathbf{r}_{\mathrm{B}}|}\right).$$

The electrostatic potential ϕ in the plane is given by $\phi = \Lambda/\sigma_{xx}$ resulting in

$$\phi(\mathbf{r}) = -\frac{I_{\mathrm{AB}}}{2\pi\sigma_{xx}}\ln\left(\frac{|\mathbf{r}-\mathbf{r}_{\mathrm{A}}|}{|\mathbf{r}-\mathbf{r}_{\mathrm{B}}|}\right). \tag{10.23}$$

The potential diverges for $\mathbf{r} \to \mathbf{r}_{\mathrm{A}}$, and for $\mathbf{r} \to \mathbf{r}_{\mathrm{B}}$. In the vicinity of one of the contacts, A or B, the equipotential lines are circles centered around the contact point. The potential at a small distance δr from contact A (here, small means $\delta r \ll |\mathbf{r}_{\mathrm{A}} - \mathbf{r}_{\mathrm{B}}|$) is (approximately) given by

$$\phi_{\mathrm{A}}(\delta r) = -\frac{I_{\mathrm{AB}}}{2\pi\sigma_{xx}}\ln\left(\frac{\delta r}{|\mathbf{r}_{\mathrm{A}}-\mathbf{r}_{\mathrm{B}}|}\right).$$

Instead of the point-like contact, we can therefore also choose a circular contact with radius δr having this potential. This choice will neither change the current density pattern nor the equipotentials. In the same way, for such a small circular contact at $\mathbf{r}_{\mathrm{B}}$ we obtain the potential

$$\phi_{\mathrm{B}}(\delta r) = \frac{I_{\mathrm{AB}}}{2\pi\sigma_{xx}}\ln\left(\frac{\delta r}{|\mathbf{r}_{\mathrm{A}}-\mathbf{r}_{\mathrm{B}}|}\right).$$

As a result, the electric resistance between two such contacts of identical radius $\delta r \ll |\mathbf{r}_{\mathrm{A}} - \mathbf{r}_{\mathrm{B}}|$ is given by

$$R_{\mathrm{AB}} = \frac{\phi_{\mathrm{A}}(\delta r) - \phi_{\mathrm{B}}(\delta r)}{I_{\mathrm{AB}}} = \frac{1}{\pi\sigma_{xx}}\ln\left(\frac{|\mathbf{r}_{\mathrm{A}}-\mathbf{r}_{\mathrm{B}}|}{\delta r}\right).$$

Van der Pauw method. The van der Pauw method is employed for the determination of the charge carrier density and the mobility if no well-defined Hall bar geometry is available. The method was suggested in 1958 by L.J. van der Pauw. The author considers a sample that fills a semi-infinite plane. Four contact points P, Q, R, and S are fabricated in a line along the edge as depicted in Fig. 10.14. The current I_{PQ} is injected through contact P and extracted through Q. The voltage is measured between the contacts R and S.

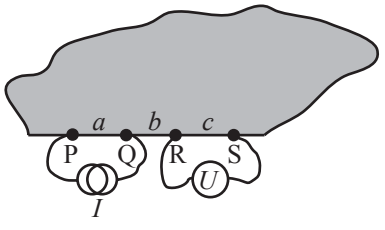


Fig. 10.14 Arrangement of four contacts along the edge of an electron gas filling a semi-infinite plane.

The electric potential difference between the two points S and R is given by

$$\Delta\phi_{\mathrm{RS}} = \phi(\mathbf{r}_{\mathrm{S}}) - \phi(\mathbf{r}_{\mathrm{R}}) = \frac{I_{\mathrm{PQ}}}{\pi\sigma_{xx}}\ln\left(\frac{|\mathbf{r}_{\mathrm{R}}-\mathbf{r}_{\mathrm{P}}||\mathbf{r}_{\mathrm{S}}-\mathbf{r}_{\mathrm{Q}}|}{|\mathbf{r}_{\mathrm{R}}-\mathbf{r}_{\mathrm{Q}}||\mathbf{r}_{\mathrm{S}}-\mathbf{r}_{\mathrm{P}}|}\right).$$

This result is obtained from eq. (10.23) by identifying the current $I_{\mathrm{AB}}/2$ with I_{PQ} due to the semi-infinite planar geometry. The four-terminal

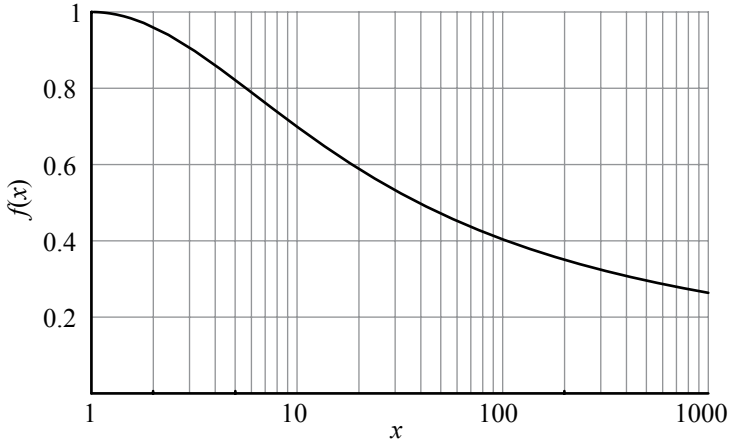


Fig. 10.15 Geometry factor for the van der Pauw method.

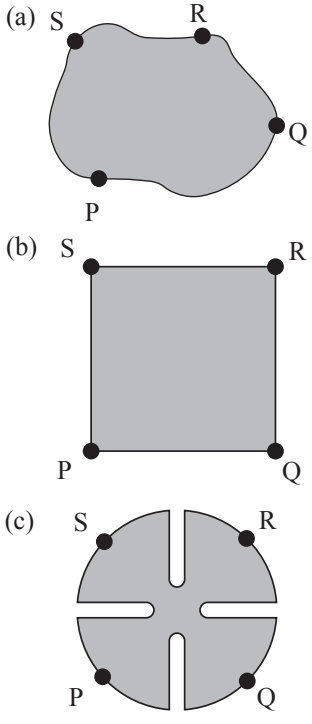


Fig. 10.16 Sample geometries for the van der Pauw method. (a) arbitrary shape (b) square shape (c) 4-foliolate sample shape.

resistance is then given by

$$R_{\mathrm{PQ,RS}} = \frac{1}{\pi\sigma_{xx}} \ln\left(\frac{|\mathbf{r}_\mathrm{R}-\mathbf{r}_\mathrm{P}||\mathbf{r}_\mathrm{S}-\mathbf{r}_\mathrm{Q}|}{|\mathbf{r}_\mathrm{R}-\mathbf{r}_\mathrm{Q}||\mathbf{r}_\mathrm{S}-\mathbf{r}_\mathrm{P}|}\right) = \frac{1}{\pi\sigma_{xx}} \ln\frac{(a+b)(b+c)}{b(a+b+c)}.$$

In a similar way we find

$$R_{\mathrm{QR,SP}} = \frac{1}{\pi\sigma_{xx}} \ln\left(\frac{|\mathbf{r}_\mathrm{S}-\mathbf{r}_\mathrm{Q}||\mathbf{r}_\mathrm{P}-\mathbf{r}_\mathrm{R}|}{|\mathbf{r}_\mathrm{S}-\mathbf{r}_\mathrm{R}||\mathbf{r}_\mathrm{P}-\mathbf{r}_\mathrm{Q}|}\right) = \frac{1}{\pi\sigma_{xx}} \ln\frac{(b+c)(a+b)}{ac}.$$

From these two equations we obtain the relation

$$e^{-\pi R_{\mathrm{PQ,RS}}\sigma_{xx}} + e^{-\pi R_{\mathrm{QR,SP}}\sigma_{xx}} = 1.$$

If the resistances $R_{\mathrm{PQ,RS}}$ and $R_{\mathrm{QR,SP}}$ are known from measurement, the conductivity σ_{xx} can be determined from this formula. For practical use we write the equation in the form

$$\rho_{xx} = \frac{1}{\sigma_{xx}} = \frac{\pi}{\ln 2}\frac{R_{\mathrm{PQ,RS}}+R_{\mathrm{QR,SP}}}{2} f\left(\frac{R_{\mathrm{PQ,RS}}}{R_{\mathrm{QR,SP}}}\right), \qquad (10.24)$$

where the function $f(x)$ depicted in Fig. 10.15 is implicitly defined by the equation

$$\frac{x-1}{x+1} = \frac{f}{\ln 2}\mathrm{acosh}\left[\frac{1}{2}\exp\left(\frac{\ln 2}{f}\right)\right].$$

Using the theory of conformal mapping, van der Pauw was able to show that eq. (10.24) remains valid for finite samples of arbitrary shape if the contacts P, Q, R, and S are sitting along the sample edge. Figure 10.16 shows examples of such samples schematically. The geometry in Fig. 10.16(c) minimizes errors arising due to the finite extent of contacts. A clever method of measuring the two resistances $R_{\mathrm{PQ,RS}}$ and $R_{\mathrm{QR,SP}}$ at the same time employs two frequencies (Kim *et al.*, 1999).

Also the Hall coefficient R_{H} can be measured in a sample of arbitrary geometry. To this end, a magnetic field B is applied normal to the plane of the electron gas, a current is driven from contact P to R, and the voltage between Q and S is measured resulting in the resistance $R_{\mathrm{PR,QS}}(B)$. Subsequently one measures $R_{\mathrm{QS,RP}}(B)$ and determines the Hall resistivity from

$$\rho_{xy}(B) = \frac{R_{PR,QS}(B) - R_{PR,QS}(0) + R_{QS,RP}(B) - R_{QS,RP}(0)}{2}.$$

10.5 Conductivity from Boltzmann's equation

Within the framework of the Drude model for the electrical conductivity described in section 10.3, the scattering time τ was introduced as a heuristic quantity. The dynamics of electrons was described classically, interactions between electrons were neglected, and scattering processes were taken into account using an ad hoc statistical average. In this section we will show how the electrical conductivity can be calculated within the framework of the Boltzmann equation. The quantum mechanical nature of electron states will be taken into account by incorporating Fermi statistics. The description of electron motion between collisions will still remain semiclassical. We will connect to elementary kinetic theory by introducing the scattering time heuristically within the relaxation time approximation. Later, in section 10.7, we extend this approach by calculating the scattering time in lowest order quantum mechanical perturbation theory.

Within the description of this section, the current density is given by

$$\mathbf{j} = -\frac{|e|}{A} \sum_{n\mathbf{k}_n\sigma} \mathbf{v}_n(\mathbf{k}_n) f_n(\mathbf{k}_n), \tag{10.25}$$

where the distribution function $f_n(\mathbf{k}_n)$ is the probability density for the occupation of state $(n\mathbf{k}_n\sigma)$, and A is a normalization area. The quantum number n labels the subband states, $\mathbf{k}$ is the wave vector of an electron in the plane of the two-dimensional electron gas, and σ is the spin quantum number. If there is no current flow through the sample, the distribution function is identical to the equilibrium Fermi–Dirac distribution. For simplicity we assume spin degeneracy and a parabolic dispersion

$$E_n(\mathbf{k}_n) = E_n + \frac{\hbar^2 k_n^2}{2m^\star}.$$

In this case the (group) velocity of an electron in subband n is given by

$$\mathbf{v}_n(\mathbf{k}_n) = \frac{\hbar \mathbf{k}_n}{m^\star},$$

and the current density becomes

$$\mathbf{j} = -\frac{2|e|\hbar}{m^\star A} \sum_{n\mathbf{k}_n} \mathbf{k}_n f_n(\mathbf{k}_n),$$

with the factor of two resulting from the spin degeneracy assumption. The electronic states are given by

$$|n\mathbf{k}_n\rangle = \frac{1}{\sqrt{A}}\chi_n(z)e^{i\mathbf{k}_n\rho},$$

where ρ describes the in-plane position of an electron. We neglect here that the wave functions are modified in a magnetic field and restrict our considerations to $\omega_c\tau \ll 1$.

The nonequilibrium distribution function $f_n(\mathbf{k}_n)$ can be determined using the Boltzmann equation (see, e.g., Ibach and Luth, 1988)

$$\frac{\partial f_n(\mathbf{k}_n)}{\partial t} + \frac{1}{\hbar}\mathbf{F}\nabla_{\mathbf{k}_n} f_n(\mathbf{k}_n) = \left(\frac{\partial f_n(\mathbf{k}_n)}{\partial t}\right)_{\text{coll}}. \tag{10.26}$$

Here we have assumed that the distribution function $f_n(\mathbf{k}_n)$ is independent of position (homogeneous electron gas). A very instructive derivation of the Boltzmann equation approach to conductivity from first principles (i.e., from the von Neuman equation of the density matrix) can be found in Kohn and Luttinger, 1957. In the following, we consider the stationary case in which $\partial f_n(\mathbf{k}_n)/\partial t = 0$. The force $\mathbf{F}$ acting on the electrons is the Lorentz force $\mathbf{F} = -|e|(\mathbf{E} + \mathbf{v} \times \mathbf{B})$.

Relaxation time approximation. For the scattering term on the right-hand side of the Boltzmann equation we use the empirical form

$$\left(\frac{\partial f_n(\mathbf{k}_n)}{\partial t}\right)_{\text{coll}} = -\frac{f_n(\mathbf{k}_n) - f_n^{(0)}(E_n(\mathbf{k}_n))}{\tau_n} \tag{10.27}$$

where we have introduced the scattering time τ_n again heuristically. If we start at time zero with some nonequilibrium distribution function, this scattering term will make sure that the distribution function returns to the equilibrium distribution within a time span of the order of τ_n. Implicit in this *Ansatz* is the assumption that electrons are not scattered between subbands. We further assume that the τ_n do not depend on the direction of $\mathbf{k}$, but they may depend on the energy via $|\mathbf{k}|$. As a consequence of the first assumption, the Boltzmann equation

$$-\frac{|e|}{\hbar}(\mathbf{E} + \frac{\hbar}{m^\star}\mathbf{k} \times \mathbf{B})\nabla_{\mathbf{k}} f(\mathbf{k}) = -\frac{f(\mathbf{k}) - f^{(0)}(E(\mathbf{k}))}{\tau} \tag{10.28}$$

can be solved separately for each subband and we have therefore dropped the subband index n. We are interested in the linear response of the distribution function to the presence of a small electric field. We linearize the Boltzmann eq. (10.28) by introducing $f(\mathbf{k}) = f^{(0)}(E(\mathbf{k})) + g(\mathbf{k})$, where $g(\mathbf{k})$ is first order in the electric field $\mathbf{E}$, and $f^{(0)}(E(\mathbf{k}))$ is the equilibrium Fermi–Dirac distribution. Inserting this expansion in eq. (10.28) and keeping only terms linear in $\mathbf{E}$ we obtain

$$-\frac{|e|\hbar}{m^\star}(\mathbf{k}\mathbf{E})\frac{\partial f^{(0)}(E(\mathbf{k}))}{dE(\mathbf{k})} - \frac{|e|}{m^\star}(\mathbf{k} \times \mathbf{B})\nabla_{\mathbf{k}} g(\mathbf{k}) = -\frac{g(\mathbf{k})}{\tau}. \tag{10.29}$$

The vector operator containing the magnetic field on the left-hand side of eq. (10.29) can be written as $(\mathbf{k}\times\mathbf{B})\nabla_{\mathbf{k}} = -\mathbf{B}(\mathbf{k}\times\nabla_{\mathbf{k}}) = -B\partial/\partial\varphi$ with φ being the angle coordinate of the wave vector $\mathbf{k}$ in cylinder coordinates. We choose $\varphi = 0$ for $\mathbf{k} \parallel \mathbf{E}$. With this result, the linearized Boltzmann eq. (10.29) reads

$$-\frac{|e|\hbar}{m^\star}(\mathbf{kE})\frac{\partial f^{(0)}(E(\mathbf{k}))}{dE(\mathbf{k})} + \omega_\mathrm{c}\frac{\partial g(\mathbf{k})}{\partial\varphi} = -\frac{g(\mathbf{k})}{\tau}. \tag{10.30}$$

This equation can be strongly simplified by introducing the function $\bar{g}(\varphi)$ related to $g(\mathbf{k})$ via

$$g(\mathbf{k}) = \frac{\partial f^{(0)}(E(\mathbf{k}))}{dE(\mathbf{k})}\hbar k\frac{|e|\tau}{m^\star}E\bar{g}(\varphi). \tag{10.31}$$

This corresponds to the idea that the equilibrium Fermi–Dirac distribution $f^{(0)}(E(\mathbf{k}))$ is at zero magnetic field shifted by a small distance $\delta\mathbf{k}$ in $\mathbf{k}$-space which corresponds to the drift velocity. With $v_\mathrm{D} = \mu\mathbf{E} = |e|\tau\mathbf{E}/m^\star = \hbar\delta\mathbf{k}/m^\star$ we find from a Taylor expansion up to first order

$$f(\mathbf{k}) = f^{(0)}(E(\mathbf{k}+\delta\mathbf{k})) = f^{(0)}(E(\mathbf{k})) + \frac{\partial f^{(0)}(E)}{\partial E}\hbar k\frac{|e|\tau\cos\varphi}{m^\star}E.$$

The second term corresponds to eq. (10.31), if $\bar{g}(\varphi) = \cos\varphi$ at zero magnetic field. Using eq. (10.31), eq. (10.30) simplifies to

$$\cos\varphi - \omega_\mathrm{c}\tau\frac{\partial\bar{g}(\varphi)}{\partial\varphi} = \bar{g}(\varphi). \tag{10.32}$$

This equation can be solved with the help of the Fourier series expansion $\bar{g}(\varphi) = \sum_\ell \bar{g}^{(\ell)}\exp(i\ell\varphi)$ leading to

$$\frac{\delta_{1\ell}+\delta_{-1\ell}}{2} - i\ell\omega_\mathrm{c}\tau\bar{g}^{(\ell)} = \bar{g}^{(\ell)}. \tag{10.33}$$

We conclude from this equation that $\bar{g}^{(\ell)} = 0$ for $|\ell| \neq 1$ and find

$$\bar{g}^{(\pm1)} = \frac{1}{2(1\pm i\omega_\mathrm{c}\tau)} = \frac{1}{2\sqrt{1+\omega_\mathrm{c}^2\tau^2}}e^{\mp i\theta} = \frac{1}{2}\cos\theta e^{\mp i\theta}, \tag{10.34}$$

where θ is the Hall angle obeying $\tan\theta = \omega_\mathrm{c}\tau$. As a consequence,

$$\bar{g}(\varphi) = \frac{\cos(\varphi-\theta)}{\sqrt{1+\omega_\mathrm{c}^2\tau^2}},$$

and

$$g(\mathbf{k}) = \frac{\partial f^{(0)}(E(\mathbf{k}))}{dE(\mathbf{k})}\hbar k\frac{|e|\tau}{m^\star}\frac{\cos(\varphi-\theta)}{\sqrt{1+\omega_\mathrm{c}^2\tau^2}}E. \tag{10.35}$$

The total nonequilibrium distribution function $f(\mathbf{k}) = f^{(0)}[E(\mathbf{k})] + g(\mathbf{k})$ is depicted in Fig. 10.17. Compared to the equilibrium distribution at $\mathbf{E} = 0$ it is shifted by an amount $\delta k = |e|\tau_0\cos\theta\,|\mathbf{E}|/\hbar$ in the direction defined by the Hall angle θ.

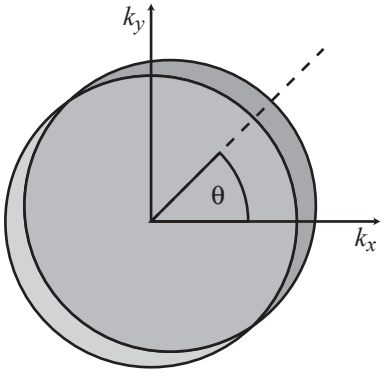


Fig. 10.17 Distribution function in k-space at finite electric and magnetic field. The distribution function is shifted from its equilibrium position by the amount $\delta k = e\tau_0\cos\theta|\mathbf{E}|/\hbar$. The direction of the shift is given by the Hall angle θ for which $\tan\theta = \omega_\mathrm{c}\tau$.

Current density and conductivity. With the above result we obtain, for the current density in eq. (10.25),

$$\mathbf{j} = \int dE \left(-\frac{\partial f^{(0)}(E)}{dE}\right) \frac{n_{\rm s}(E) e^2 \tau(E)/m^\star}{\sqrt{1+\omega_{\rm c}^2\tau^2(E)}} \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} |\mathbf{E}|,$$

where $n_{\rm s}(E)$ is the electron density obtained from occupying states up to the energy E (at zero temperature). The expressions for the conductivity tensor components are then

$$\sigma_{xx}(B,T) = \int dE \left(-\frac{\partial f^{(0)}(E)}{dE}\right) \frac{n_{\rm s}(E) e^2 \tau(E)/m^\star}{1+\omega_{\rm c}^2\tau^2(E)} \quad (10.36)$$

$$\sigma_{xy}(B,T) = \int dE \left(-\frac{\partial f^{(0)}(E)}{dE}\right) \frac{n_{\rm s}(E) e^2 \omega_{\rm c}\tau^2(E)/m^\star}{1+\omega_{\rm c}^2\tau^2(E)} \quad (10.37)$$

Equations (10.36) and (10.37) have the structure

$$\sigma_{xx}(B,T) = \int dE \left(-\frac{\partial f^{(0)}(E)}{dE}\right) \sigma_{xx}(B,E,T=0) \quad (10.38)$$

$$\sigma_{xy}(B,T) = \int dE \left(-\frac{\partial f^{(0)}(E)}{dE}\right) \sigma_{xy}(B,E,T=0), \quad (10.39)$$

emphasizing that the finite temperature conductivity can be obtained from a zero temperature energy-dependent conductivity. However, this formulation has to be interpreted with great care because the energy-dependent scattering rate $\tau^{-1}(E)$ can implicitly depend on the temperature via the temperature dependence of impurity potential screening, as we will see later.

An important property of these results is the derivative of the Fermi function in the integrand of the energy integral. At low temperatures, for which $k_{\rm B}T \ll E_{\rm F}$, this derivative has a very sharp maximum at $E = E_{\rm F}$, and the components of the conductivity tensor become

$$\sigma_{xx}(B,E_{\rm F},T=0) = \frac{n_{\rm s}e^2\tau(E_{\rm F})}{m^\star} \frac{1}{1+\omega_{\rm c}^2\tau^2(E_{\rm F})}$$

$$\sigma_{xy}(B,E_{\rm F},T=0) = \frac{n_{\rm s}e^2\tau(E_{\rm F})}{m^\star} \frac{\omega_{\rm c}\tau(E_{\rm F})}{1+\omega_{\rm c}^2\tau^2(E_{\rm F})}$$

in agreement with eqs. (10.12) and (10.13). Compared to the much simpler derivation presented earlier we have now learned that, as a consequence of the Pauli principle introduced via the Fermi–Dirac distribution function, the scattering time τ has to be evaluated at the Fermi energy. *The low-temperature conductivity therefore reflects the scattering properties of the electron gas at the Fermi edge.*

As a final step we determine the components of the specific resistivity for small magnetic fields, i.e., $\omega_{\rm c}\tau \ll 1$. To this end, we define the average powers of the scattering time

$$\langle \tau^n \rangle = \int dE \left(-\frac{\partial f^{(0)}(E)}{dE}\right) \frac{E}{E_{\rm F}} \tau^n(E). \quad (10.40)$$

The elements of the conductivity tensor are then, to first order in B, given by

$$\begin{aligned}\sigma_{xx}(B,T) &= \frac{n_{\mathrm{s}}e^2\langle\tau\rangle}{m^\star}\\ \sigma_{xy}(B,T) &= \frac{n_{\mathrm{s}}e^2\omega_{\mathrm{c}}\langle\tau^2\rangle}{m^\star},\end{aligned}$$

and the resulting specific resistivities for $B \to 0$ are

$$\begin{aligned}\rho_{xx}(B,T) &= \frac{m^\star}{n_{\mathrm{s}}e^2\langle\tau\rangle}\\ \rho_{xy}(B,T) &= \frac{B}{n_{\mathrm{s}}|e|}\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2} := \frac{r_{\mathrm{H}}(T)B}{n_{\mathrm{s}}|e|}.\end{aligned}$$

The expression for the Hall resistivity differs from eq. (10.16) in the appearance of the temperature-dependent factor $r_{\mathrm{H}}(T) = \langle\tau^2\rangle / \langle\tau\rangle^2$. It turns out that for sufficiently low temperatures this factor is one and the Drude expressions in eqs. (10.15) and (10.16) can be used for the determination of the charge carrier density and the scattering time. For higher temperatures $r_{\mathrm{H}}(T)$ remains of the order of one.

Mean free path. We are now going to define the important length scale for elastic impurity scattering in the diffusive transport regime. Because the relevant electron scattering time is evaluated at the Fermi energy at low temperatures, the mean free path can be defined as

$$l = v_{\mathrm{F}}\tau_{\mathrm{F}}. \tag{10.41}$$

This length scale can be compared to other length scales such as the mean electron separation or the characteristic size of the nanostructure under consideration. With this definition, the Drude conductivity in eq. (10.12) evaluated at $B = 0$ for two-dimensional electron gases in a spin degenerate conduction band minimum can be expressed as

$$\sigma = \frac{e^2}{h}k_{\mathrm{F}}l.$$

The product $k_{\mathrm{F}}l$ is a measure for the ability of scatterers (or the spatially fluctuating potential) to localize electrons. For $k_{\mathrm{F}}l \gg 1$ the tendency for electron localization is weak and one talks about metal-like conduction. For $k_{\mathrm{F}}l \ll 1$ electrons localize strongly in the potential minima of the fluctuating potential. The quantity $e^2/h \approx (26\,\mathrm{k\Omega})^{-1}$ is the so-called conductance quantum.

10.6 Scattering mechanisms

In two-dimensional electron gases, the most important scattering mechanisms are

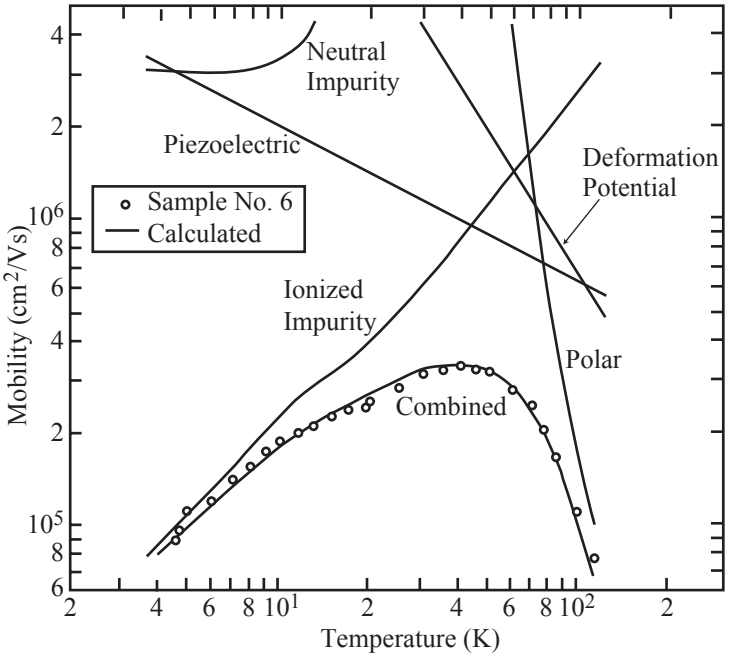


Fig. 10.18 Influence of various scattering mechanisms on the temperature dependence of the mobility of a three-dimensional GaAs sample. (Reprinted from Sequoia *et al.*, 1976 with permission from Elsevier.)

- optical phonon scattering (dominant at high temperatures)
- acoustic phonon scattering (deformation potential scattering)
- piezoelectric scattering originating from acoustic phonons in piezoelectric semiconductors (e.g., III-V semiconductors or II-VI semiconductors, as a result of the lack of bulk inversion symmetry of the crystal lattice)
- ionized impurity scattering (undesired background doping)
- ionized donor scattering
- scattering from neutral defects or impurities
- alloy scattering in ternary semiconductors, e.g., in AlGaAs heterostructures
- surface roughness scattering.

All the above scattering mechanisms can lead to intersubband scattering, if more than a single subband is occupied. Figure 10.18 shows the influence of various scattering mechanisms on the temperature-dependent mobility of a three-dimensional GaAs sample as a reference. At temperatures above 100 K polar optical phonon scattering is by far the dominant scattering mechanism limiting the mobility. In the intermediate range between 40 K and 100 K, various scattering mechanisms play a role. Phonon scattering dies out when the temperature is lowered, but ionized impurity scattering becomes more and more dominant. Below about 10 K, ionized impurity scattering appears to be the only relevant scattering mechanism.

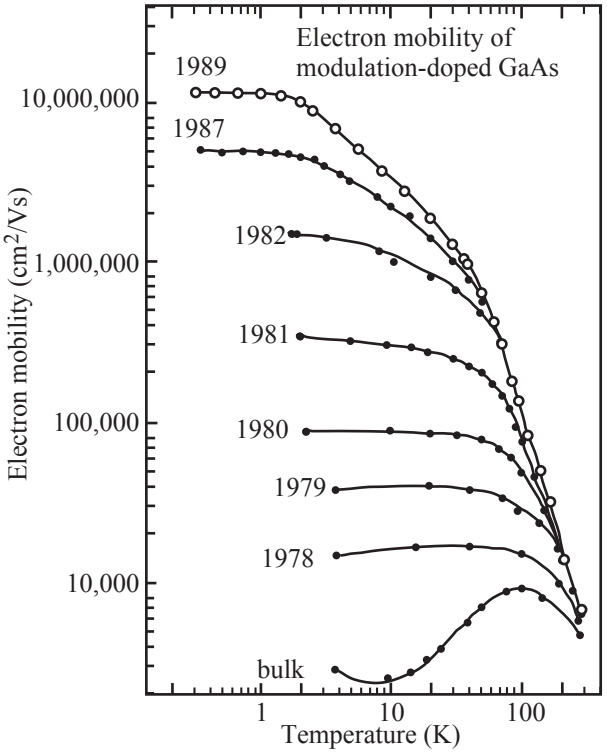


Fig. 10.19 Improvement of the low-temperature mobilities in remotely doped Ga[Al]As heterostructures. (Reprinted with permission from Pfeiffer *et al.*, 1989. Copyright 1989, American Institute of Physics.)

The situation is different in some respects, if one compares the three-dimensional case to that of two-dimensional electron gases. Figure 10.19 shows how the experimentally achievable temperature dependence of the mobility has changed over the years since 1978 with steady improvement in growth techniques and therefore in material quality. The mobilities of three-dimensional GaAs samples cannot compete with those of two-dimensional systems as a result of the remote doping technique. Figure 10.20(a) shows which scattering mechanisms are relevant for the mobility of electrons in an optimized two-dimensional gas in a Ga[Al]As heterostructure. As in bulk GaAs, the mobility is mainly determined by optical phonon scattering at temperatures above 100 K. At intermediate temperatures between 20 K and 40 K, various scattering mechanisms contribute, as they do in the three-dimensional system. However, ionized impurity scattering from remote impurities (dopants) has very little influence on the mobility owing to the separation of the electron gas from the doping plane. It is rather the background impurity doping which appears to limit the low-temperature mobility. Therefore, improved growth techniques in dedicated MBE systems which lead to an improved material quality in terms of lower background impurity levels have enabled us to achieve mobilities of more than $10^7\,\mathrm{cm}^2/\mathrm{Vs}$ (see Fig. 10.19).

The mobility of electrons in a heterostructure with remote doping is crucially influenced by the thickness of the spacer layer between the electron gas and the two-dimensional electron gas. Figure 10.20(b) shows

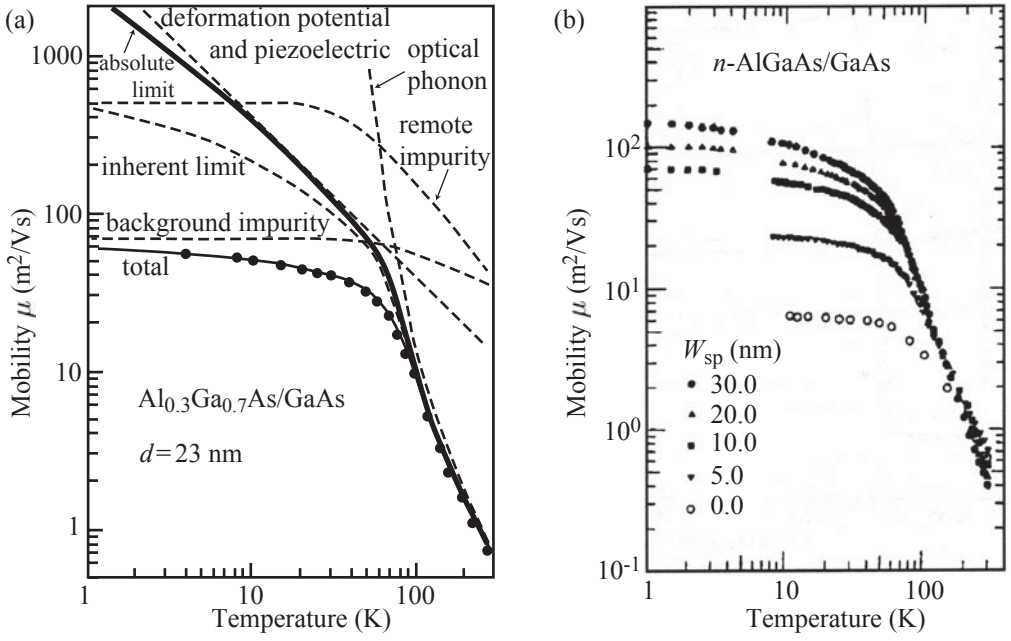


Fig. 10.20 (a) Influence of various scattering mechanisms on the mobility of a GaAs/AlGaAs heterostructure. (Reprinted with permission from Walukiewicz *et al.*, 1984. Copyright 1984 by the American Physical Society.) (b) Mobilities of electrons in modulation doped GaAs/AlGaAs heterostructures with varying spacer layer thickness W_{sp} (Solomon *et al.*, 1984).

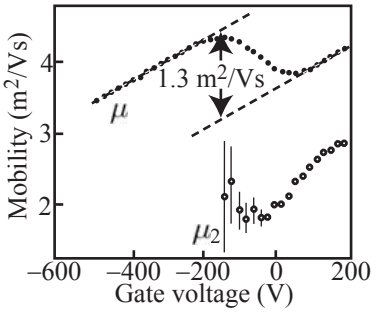


Fig. 10.21 Mobility of a GaAs/AlGaAs heterostructure as a function of the top gate voltage. Above about −50 mV the second subband is occupied. Additional intersubband scattering leads to a reduction of the mobility. (Reprinted from Stormer *et al.* 1982 with permission from Elsevier.)

the mobilities of samples with different spacer layer thicknesses W_{sp} between zero and 30 nm. The influence of E_{sp} is quite significant below about 50 K. The biggest improvement of the low-temperature mobility is achieved when W_{sp} is increased from zero to 10 nm. This increases the mobility by more than one order of magnitude. Beyond $W_{\mathrm{sp}} = 10$ nm the mobility increase becomes much weaker, because background impurity scattering becomes more and more dominant.

The influence of intersubband scattering can be made visible in experiments in which the number of occupied subbands can be controlled. One option is the use of a top gate voltage which may allow us to increase the electron density beyond the occupation threshold of the second subband. Another method uses electron gases in which two subbands are already occupied at zero magnetic field. The higher subband can then be depopulated by applying a magnetic field in parallel to the electron gas. Fig. 10.21 shows the resistivity of a modulation-doped heterostructure in which the the second subband is populated at a top gate voltage of about −50 mV. The mobility decreases as a result of the intersubband scattering channel that becomes important above this gate voltage.

10.7 Quantum treatment of ionized impurity scattering

In section 10.5 we treated scattering on a heuristic basis by introducing the relaxation times τ_n for each subband n. The resulting expressions for the conductivity tensor components in eq. (10.36) and (10.37) are compatible with the earlier Drude results, but it remained unclear how the actual scattering times τ_n have to be calculated from microscopic scattering theory. We will study the basic ideas behind the microscopic determination of the relaxation rates below, using lowest order scattering at ionized impurities as an example. Higher order contributions, as well as interference effects, are therefore neglected, but intersubband scattering processes are taken into account. A very rigorous approach to this problem starting from first principles is given in Kohn and Luttinger 1957.

The basic ingredient in our approach is the introduction of appropriate quantum scattering rates in the collision term of the Boltzmann eq. (10.26), instead of the empirical expression (10.27):

$$\left(\frac{\partial f_n(\mathbf{k}_n)}{\partial t}\right)_{\text{coll}} = \sum_{m\mathbf{k}'_m} \{W_{mn}(\mathbf{k}'_m, \mathbf{k}_n)[1 - f_n(\mathbf{k}_n)]f_m(\mathbf{k}'_m) - W_{nm}(\mathbf{k}_n, \mathbf{k}'_m)[1 - f_m(\mathbf{k}'_m)]f_n(\mathbf{k}_n)\} \quad (10.42)$$

Here, the $W_{nm}(\mathbf{k}_n, \mathbf{k}'_m)$ are the quantum scattering rates from state $(n\mathbf{k}_n)$ into state $(m\mathbf{k}'_m)$. In eq. (10.42) the first term in curly brackets describes scattering processes from any state $(m\mathbf{k}'_m)$ *into* the state $(n\mathbf{k})$. This process is only possible if the initial state is occupied (factor $f_m(\mathbf{k}'_m)$), and if the final state is unoccupied (factor $1 - f_n(\mathbf{k}_n)$). The second term describes scattering *out* of the state $(n\mathbf{k})$ into any other state $(m\mathbf{k}'_m)$.

Calculation of the elastic scattering rates. The scattering rates are calculated in first order perturbation theory using Fermi's golden rule

$$W_{nm}(\mathbf{k}_n, \mathbf{k}'_m) = \frac{2\pi}{\hbar} \left|\langle m\mathbf{k}'_m \left|V\right| n\mathbf{k}_n\rangle\right|^2 \delta(E_n(\mathbf{k}_n) - E_m(\mathbf{k}'_m)).$$

Here we are interested in elastic scattering processes that are dominant at low temperatures where phonons are essentially frozen out. For these processes we have the scattering matrix elements

$$\begin{aligned}\langle m\mathbf{k}'_m \left|V\right| n\mathbf{k}_n\rangle &= \frac{1}{A}\int dz d^2\rho\, \chi_m(z) e^{-i\mathbf{k}'_m\rho} V(\rho, z)\chi_n(z) e^{i\mathbf{k}_n\rho} \\ &= \frac{1}{A}\int dz d^2\rho\, \chi_m(z)\chi_n(z) e^{-i(\mathbf{k}'_m - \mathbf{k}_n)\rho} V(\rho, z) \\ &:= \frac{1}{A} V_{mn}(\mathbf{k}'_m - \mathbf{k}_n).\end{aligned}$$

The plane-wave wave functions in the plane of the electron gas essentially lead to a two-dimensional Fourier transform of the scattering potential

in the plane. Diagonal matrix elements $V_{nn}(\mathbf{k}'_n - \mathbf{k}_n)$ describe scattering within subband n (*intra*subband scattering), whereas off-diagonal matrix elements with $n \neq m$ describe intersubband scattering between subbands n and m. We further assume that the scattering potential $V(\rho, z)$ is the sum over individual scattering centers with potentials $v_i(\rho, z)$ localized at positions ρ_i, i.e.,

$$V(\rho, z) = \sum_i v_i(\rho - \rho_i, z).$$

With this assumption, the scattering matrix element can be written as

$$V_{nm}(\mathbf{q}) = \sum_i v_{nm}^{(i)}(\mathbf{q}) e^{i\mathbf{q}\rho_i},$$

where we have introduced the change in wave vector $\mathbf{q} = \mathbf{k}'_m - \mathbf{k}_n$. The modulus squared of the matrix element is then given by

$$\begin{aligned} \left|\langle m\mathbf{k}'_m \left|V\right| n\mathbf{k}_n\rangle\right|^2 &= \frac{1}{A^2} \sum_{ij} v_{nm}^{(i)\star}(\mathbf{q}) v_{nm}^{(j)}(\mathbf{q}) e^{i\mathbf{q}(\rho_j - \rho_i)} \\ &= \frac{1}{A^2} \sum_i \left|v_{nm}^{(i)}(\mathbf{q})\right|^2 \\ &\quad + \frac{1}{A^2} \sum_{ij, i\neq j} v_{nm}^{(i)\star}(\mathbf{q}) v_{nm}^{(j)}(\mathbf{q}) e^{i\mathbf{q}(\rho_j - \rho_i)}. \end{aligned}$$

In the last step, we have separated the diagonal contributions $i = j$ from the off-diagonal contributions $i \neq j$. If we assume that the scattering centers are randomly placed, i.e., the pair correlation function is constant, the statistical phases in the exponential of the off-diagonal term will mutually cancel and the term vanishes. The remaining diagonal term can be written as

$$\left|\langle m\mathbf{k}_m \left|V\right| n\mathbf{k}_n\rangle\right|^2 = \frac{N_i}{A^2} \frac{1}{N_\mathrm{i}} \sum_i \left|v_{nm}^{(i)}(\mathbf{q})\right|^2 := \frac{N_i}{A^2} \overline{\left|v_{nm}^{(i)}(\mathbf{q})\right|^2},$$

where N_i is the number of scattering centers within the normalization area A. We have introduced the averaged Fourier transform $\overline{\left|v_{nm}^{(i)}(\mathbf{q})\right|^2}$ of the scattering potentials. This average will, of course, depend on the specific locations ρ_i of the N_i scattering centers. However, it has been argued (Kohn and Luttinger, 1957) that, for the purpose of calculating the scattering rate, an average of this quantity taken over the ensemble of all possible impurity configurations can be used equivalently, such that the squared matrix element becomes *independent* of the specific impurity configuration in a sample. Denoting this impurity average by $\left\langle \left|v_{nm}^{(i)}(\mathbf{q})\right|^2 \right\rangle_\mathrm{imp}$, we get for the squared matrix element the expression

$$\left\langle \left|\langle m\mathbf{k}_m \left|V\right| n\mathbf{k}_n\rangle\right|^2 \right\rangle_\mathrm{imp} = \frac{N_i}{A^2} \left\langle \left|v_{nm}^{(i)}(\mathbf{q})\right|^2 \right\rangle_\mathrm{imp}.$$

This function will often not depend on the direction, but only on the magnitude of $\mathbf{q}$. In this case we find

$$|\mathbf{q}| = \sqrt{(\mathbf{k}_n - \mathbf{k}'_m)^2} = \sqrt{{k'_m}^2 + k_n^2 - 2k'_m k_n \cos(\varphi_n - \varphi'_m)},$$

i.e., the average scattering potential depends only on the energy and the scattering angle $\varphi_n - \varphi'_m$. We therefore define

$$\left\langle \left| v^{(i)}_{nm}(\mathbf{q}) \right|^2 \right\rangle_{\text{imp}} := P_{nm}[E_n(\mathbf{k}_n), |\varphi_n - \varphi'_m|].$$

After all these considerations, assumptions, and simplifications we can write the scattering rate as

$$W_{nm}(\mathbf{k}_n - \mathbf{k}'_m) = \frac{2\pi}{\hbar}\frac{N_i}{A^2} P_{nm}[E_n(\mathbf{k}_n), |\varphi_n - \varphi'_m|]\delta(E_n(\mathbf{k}_n) - E_m(\mathbf{k}'_m)). \quad (10.43)$$

Linearized Boltzmann equation. Taking into account the symmetry $W_{nm}(\mathbf{k}_n - \mathbf{k}'_m) = W_{mn}(\mathbf{k}'_m - \mathbf{k}_n)$ evident from eq. (10.43), the scattering term, eq. (10.42), simplifies, and we can write the linearized Boltzmann equation in complete analogy with eq. (10.30) as

$$-\frac{|e|\hbar}{m^\star}(\mathbf{k}_n\mathbf{E})\frac{\partial f^{(0)}(E_n(\mathbf{k}_n))}{dE_n(\mathbf{k}_n)} + \omega_c \frac{\partial g_n(\mathbf{k}_n)}{\partial \varphi_n} = \sum_{m\mathbf{k}_m} W_{mn}(\mathbf{k}_m - \mathbf{k}_n)\left\{g_m(\mathbf{k}_m) - g_n(\mathbf{k}_n)\right\}. \quad (10.44)$$

For obtaining the right-hand side, we have used the fact that the scattering processes are elastic, i.e., $E_n(\mathbf{k}_n) = E_m(\mathbf{k}'_m)$.

By analogy with eq. (10.31) we now define

$$g_n(\mathbf{k}_n) := \frac{\partial f^{(0)}(E_n(\mathbf{k}_n))}{dE_n(\mathbf{k}_n)}\hbar k_n \frac{|e|\tau_n(\varphi_n)}{m^\star}E, \quad (10.45)$$

where the product $\tau\bar{g}(\varphi)$ appearing in eq. (10.31) has been contracted into the angle-dependent scattering time $\tau_n(\varphi)$. Using eqs. (10.45) and (10.43), eq. (10.44) becomes

$$k_n \cos\varphi_n - k_n\omega_c \frac{\partial \tau_n(\varphi_n)}{\partial \varphi_n} = \frac{n_i m^\star}{2\pi\hbar^3} \sum_m \int_0^{2\pi} d\varphi'_m P_{mn}[E_n(\mathbf{k}_n), |\varphi'_m - \varphi_n|]\left\{k_n\tau_n(\varphi_n) - k'_m\tau_m(\varphi'_m)\right\}, \quad (10.46)$$

where we have introduced the areal density of scatterers $n_\text{i} := N_\text{i}/A$. This equation can be solved with the help of the Fourier series expansions $\tau_n(\varphi_n) = \sum_\ell \tau_n^{(\ell)} \exp(i\ell\varphi_n)$ and $P_{mn}(E,\varphi) = \sum_\ell P_{mn}^{(\ell)}(E)\exp(i\ell\varphi)$

which leads to

$$k_n\left[\frac{\delta_{1j}+\delta_{-1j}}{2}-i\omega_{\rm c}\tau_n^{(j)}j\right] = \frac{n_i m^\star}{\hbar^3}\sum_m\left\{P_{mn}^{(0)}[E_n(\mathbf{k}_n)]k_n\tau_n^{(j)}-P_{mn}^{(j)}[E_n(\mathbf{k}_n)]k'_m\tau_m^{(j)}\right\}, \quad (10.47)$$

which corresponds to eq. (10.33) in our previous derivation of the distribution function. We conclude from this equation that $\tau_n^{(j)}=0$ for $|j|\neq 1$ and are therefore left with the equation

$$\frac{1}{2}k_n = \sum_m\left\{\delta_{mn}\left[\left(\sum_\ell\frac{n_i m^\star}{\hbar^3}P_{\ell n}^{(0)}[E_n(\mathbf{k}_n)]\right)\pm i\omega_{\rm c}\right] \quad (10.48)$$

$$-\frac{n_i m^\star}{\hbar^3}P_{mn}^{(\pm1)}[E_n(\mathbf{k}_n)]\right\}k_m\tau_m^{(\pm1)}. \quad (10.49)$$

Scattering rate in the single subband case. In the case of a single occupied subband (quantum limit) we obtain

$$\frac{1}{2}=\left\{\frac{n_i m^\star}{\hbar^3}P^{(0)}(E)\pm i\omega_{\rm c}-\frac{n_i m^\star}{\hbar^3}P^{(1)}(E)\right\}\tau^{(\pm1)}$$

with the solution [cf., eq. (10.34)]

$$\tau^{(\pm1)}=\frac{1}{2}\frac{\tau_0}{1\pm i\omega_{\rm c}\tau_0}=\frac{1}{2}\frac{\tau_0 e^{\mp i\theta}}{\sqrt{1+\omega_{\rm c}^2\tau_0^2}}=\frac{1}{2}\tau_0\cos\theta e^{\mp i\theta},$$

where θ is the Hall angle and the zero magnetic field scattering rate is given by

$$\frac{\hbar}{\tau_0(E)}=n_i\frac{m^\star}{2\pi\hbar^2}\int_0^{2\pi}d\varphi\left\langle\left|v^{(i)}(\mathbf{q})\right|^2\right\rangle_{\rm imp}(1-\cos\varphi). \quad (10.50)$$

The prefactor $m^\star/2\pi\hbar^2$ in front of the integral is half of the two-dimensional density of states (half, because scattering conserves spin and only half of the total density of states is therefore available for scattering into). The nonequilibrium part of the distribution function is identical to eq. (10.35) with $\tau\to\tau_0(E)$. Equation (10.50) is our main result, an expression for the calculation of the energy-dependent Drude scattering rate from microscopic scattering theory. It can be used for the calculation of the conductivity tensor components in eqs. (10.36) and (10.37) by replacing $\tau(E)\to\tau_0(E)$. The scattering potential matrix element $v^{(i)}(\mathbf{q})$ will typically be the result of a screened ionized impurity potential. Using the results of linear screening theory, eq. (9.12), we can therefore express the scattering rate as

$$\frac{\hbar}{\tau_0(E)}=n_i\frac{m^\star}{2\pi\hbar^2}\int_0^{2\pi}d\varphi\left\langle\frac{\left|v^{(\rm ext)}(\mathbf{q})\right|^2}{\epsilon^2(q,E_{\rm F},T)}\right\rangle_{\rm imp}(1-\cos\varphi). \quad (10.51)$$

The temperature and density dependence of the dielectric function adds an interaction-related temperature and density dependence to the conductivity in eqs. (10.36) and (10.37) beyond the derivative of the Fermi distribution function.

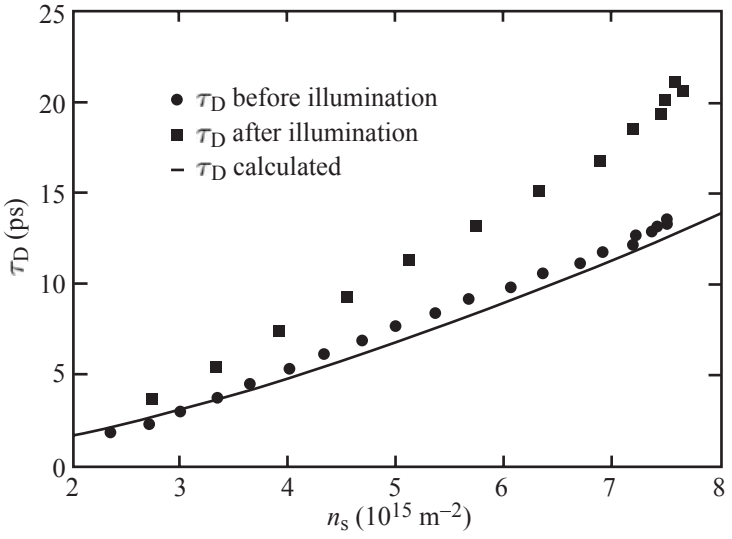


Fig. 10.22 Density dependence of the Drude scattering time in a 10 nm thick GaAs quantum well at $T = 1.7$ K. The symbols are data points of the sample before and after illumination with light from an infrared LED. The solid line is the result of a model calculation.

Electron density dependence of the scattering time. At low temperatures, where ionized background impurity scattering is dominant, the average scattering time $\langle\tau_0\rangle \approx \tau_0(E_F)$ depends on the electron density. Figure 10.22 shows this dependence as it is observed in a 10 nm thick GaAs quantum well at a temperature $T = 1.7$ K. The scattering time increases with increasing electron density. This behavior can be explained with eq. (10.50). In the present case of a single occupied subband

$$|\mathbf{q}| = \sqrt{2k_F^2(1-\cos\varphi)} = \sqrt{4\pi n_s(1-\cos\varphi)}.$$

The average Fourier transform of the scattering potential $\left\langle \left|v^{(i)}(\mathbf{q})\right|^2 \right\rangle_{\text{imp}}$ will have a maximum for $|\mathbf{q}| \to 0$ and decrease monotonically for increasing q as shown schematically in Fig. 10.23. The angle integration in eq. (10.50) averages over q-values between 0 and $2k_F$. Scattering angles close to π, i.e., large q-values (backscattering) get a strong weight. The value of this angular average decreases with increasing density, i.e., growing k_F, because $\left\langle \left|v^{(i)}(2k_F)\right|^2 \right\rangle_{\text{imp}}$ decreases. As a result, the scattering rate decreases and the scattering time increases.

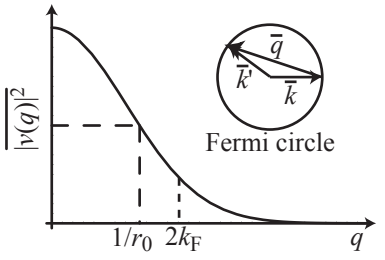


Fig. 10.23 The squared modulus of a typical scattering matrix element. At the top right, scattering vectors for a particular scattering process are shown. The maximum possible value for q is $2k_F$.

10.8 Einstein relation: conductivity and diffusion constant

In a phenomenological approach to the conductance, valid in one ($d = 1$), two ($d = 2$), or three $d = 3$ dimensions, we can regard the total current in a conductor as the sum of a so-called drift current driven by the electric field [cf., eq. (10.2)]

$$\mathbf{j}_{\text{Drift}} = \sigma\mathbf{E}$$

and a diffusion current caused by a gradient in the electron density

$$\mathbf{j}_{\mathrm{Diff}} = |e| D_d \nabla n.$$

Here, D is the diffusion constant in a d-dimensional system (note that the diffusion constant has dimensions m^2/s in $d = 1, 2, 3$).

In thermodynamic equilibrium, the electrochemical potential μ_{elch} is constant, and the total current in a conductor is zero, i.e.,

$$\sigma \mathbf{E} + |e| D_d \nabla n_{\mathrm{s}} = 0 \text{ for } \nabla \mu_{\mathrm{elch}} = 0. \tag{10.52}$$

The electrochemical potential is the sum of the chemical potential E_{F} (i.e., the Fermi energy measured from the minimum of the dispersion relation), and of the electrostatic potential $-|e|\phi$, i.e., $\mu_{\mathrm{elch}} = E_{\mathrm{F}} - |e|\phi$. As a result we obtain, for zero temperature,

$$\nabla \mu_{\mathrm{elch}} = \nabla E_{\mathrm{F}} + |e| \mathbf{E} = \frac{1}{\mathcal{D}_d(E_{\mathrm{F}})} \nabla n_{\mathrm{s}} + |e| \mathbf{E}. \tag{10.53}$$

Here, $\mathcal{D}_d(E_{\mathrm{F}}) = dn_{\mathrm{s}}(E_{\mathrm{F}})/dE_{\mathrm{F}}$ is the density of states at the Fermi energy in a d-dimensional system. If we combine eqs. (10.52) and (10.53) we obtain the so-called Einstein relation for an electron gas at zero temperature

$$\sigma = e^2 \mathcal{D}_d(E_{\mathrm{F}}) D_d, \tag{10.54}$$

which is a relation between the electric transport problem, characterized by the conductance σ, and the diffusion problem, characterized by the diffusion constant D_d. Comparison with eq. (10.12) for $B = 0$ leads to the zero magnetic field expression for the diffusion constant in two dimensions ($d = 2$)

$$D_2 = \frac{1}{2} v_{\mathrm{F}}^2 \tau = \frac{1}{2} \frac{l^2}{\tau}. \tag{10.55}$$

At finite magnetic fields, the diffusion constant in two dimensions, like the conductivity, becomes a 2×2 tensor. For $\omega_{\mathrm{c}} \tau \gg 1$ we find, with the help of σ_{xx} and the Einstein relation,

$$D_{xx} = \frac{1}{2} \frac{R_{\mathrm{c}}^2}{\tau}, \tag{10.56}$$

where $R_{\mathrm{c}} = \hbar k_{\mathrm{F}}/eB$ is the classical cyclotron radius. In a strong magnetic field ($R_{\mathrm{c}} \ll l$, i.e., $\omega_{\mathrm{c}} \tau \gg 1$), this length scale takes the role of the mean free path for the electronic motion. Intuitively, this result makes sense if one realizes that a scattering event will make the center coordinate of the cyclotron radius jump by some distance between zero and $2R_{\mathrm{c}}$, i.e., by R_{c} on average. According to this point of view, the conductance at strong magnetic fields is determined by the diffusion of the center coordinates of the cyclotron orbits.

10.9 Scattering time and cross-section

Within the phenomenological scattering theory, the scattering time has a simple relation to the scattering cross-section. As an example, we

consider scattering of electrons at lattice defects. The differential cross-section $\sigma_s(\Omega)$ describes scattering of an electron at a single lattice defect. If F is the current of the incident electrons and dN is the number of electrons scattered into the solid angle $d\Omega$ around Ω, then

$$dN = F\sigma_s(\Omega)\, d\Omega.$$

The total scattering cross-section σ_{tot} is obtained by integrating over all solid angles. Intuitively, the total cross-section is the effective area of the scattering center. If a particle hits this area, it is scattered, otherwise it is not. If an electron moves with velocity v in an electron gas in which scatterers exist with a density N_i, the probability for an electron to scatter within time interval dt is given by

$$\frac{dt}{\tau} = N_i \sigma_s v dt.$$

The relation between total scattering cross-section and scattering rate is therefore

$$\frac{1}{\tau} = N_i \sigma_s v, \tag{10.57}$$

where N_i is the density of defects and v is the average velocity of electrons. The scattering cross-section σ_s and the scattering rate $1/\tau$ will in general depend on the energy of the particle under consideration. Of crucial importance for the conductivity of two-dimensional electron gases at low temperatures is the scattering rate at the Fermi energy $\tau_F^{-1} = N_i \sigma_s(E_F) v_F$.

10.10 Conductivity and field effect in graphene

At the end of this chapter on classical Drude transport we briefly discuss the conductivity and the field effect in two-dimensional graphene. A measurement of the conductivity of graphene performed at a temperature of 1.7 K is shown in Fig. 10.24. The measurement was performed on a Hall bar structure in four-terminal configuration. The graphene flake was deposited on a highly doped silicon substrate acting as a back gate. An oxide barrier of $d = 300$ nm thickness separated this back gate from the graphene Hall bar which was patterned with lithographic techniques.

The influence of the back gate on the graphene sheet can be described using a parallel plate capacitor model with

$$\Delta n_s(V_{bg}) = \frac{\epsilon\epsilon_0}{|e|d} \Delta V_{bg}$$

describing the relation between the change in sheet carrier density Δn_s and the change in back gate voltage ΔV_{bg}. Here, ε is the relative dielectric constant of the oxide. In the measurement, the conductivity is seen to increase almost linearly to the left and to the right of the so-called charge neutrality point V_{bg}^D, as the above formula suggests for the

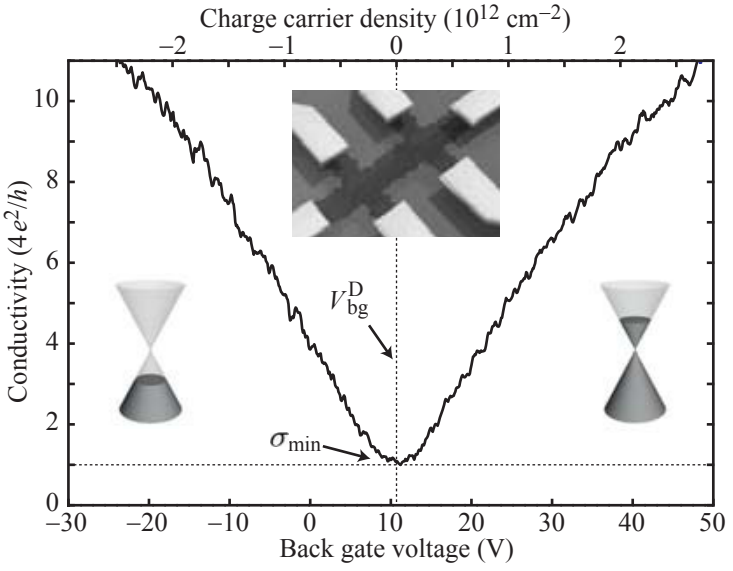


Fig. 10.24 Conductivity of graphene measured on a Hall bar structure (see inset for a schematic drawing) as a function of the back gate voltage.

density. At gate voltages $V_{bg} < V_{bg}^{D}$, the Fermi energy is in the valence band, whereas in the opposite case it is in the conduction band. This is schematically indicated with the dispersion cones in the figure. The conductivity σ resembles the v-shaped density of states in graphene (see Fig. 3.15) except that the conductivity does not vanish at V_{bg}^{D}, but has a minimum σ_{min} at a value of about $4e^2/h = 0.15 \times 10^{-3}\,\Omega^{-1}$. We anticipate here that the prefactor of four can be seen as a result of the two-fold spin degeneracy and the two-fold valley degeneracy (K and K′) in graphene. The remaining value of e^2/h is called the conductance quantum, which is of fundamental importance for quantum transport, as we will see in later chapters.

Using the Drude–Boltzmann expression $\sigma = n_s|e|\mu$ for the conductivity at zero magnetic field, we have to conclude that the mobility μ in graphene is almost independent of the charge carrier density n_s. An interesting aspect of the mobility of charge carriers in graphene is the fact that the linear dispersion relation of graphene at the Fermi energy does not allow us to define an effective mass. As a consequence, the relation $\mu = |e|\tau/m^\star$, which is valid in the case of parabolic dispersion relations, does not apply here. We obtain a different view on the problem if we consider the Einstein relation (10.54). The linear dependence of the conductivity to the left and right of V_{bg}^{D} implies that the diffusion constant in graphene is also independent on the energy.

A detailed analysis of the Drude–Boltzmann theory of conductivity for graphene, with its peculiar band structure, reveals more insights. As a consequence of the two-component wave function describing the charge carriers in graphene, the scattering matrix elements produce an additional scattering-angle-dependent factor suppressing backscattering of carriers (scattering angles of π). The scattering rate can be written

as [cf., eq. (10.50)]

$$\frac{\hbar}{\tau_0(E)} = n_\mathrm{i}\frac{E/c^{\star 2}}{2\pi\hbar^2}\int_0^{2\pi} d\varphi \left\langle \left| v^{(i)}(\mathbf{q}) \right|^2 \right\rangle_\mathrm{imp} \frac{1+\cos\varphi}{2}(1-\cos\varphi).$$

Backscattering is suppressed ($1+\cos\varphi$ factor), because the two-component states with wave vectors $\mathbf{k}$ and $-\mathbf{k}$ are orthogonal (they have the opposite helicity, but we have assumed that scattering conserves the helicity). We also see that the mass in eq. (10.50) has to be replaced by the relativistic mass $E/c^{\star 2}$ for the case of graphene. Furthermore, the prefactor in front of the integral resembles one quarter of the graphene density of states in eq. (3.27). It is only one quarter of the full density of states because scattering conserves the spin, and intervalley scattering is not admitted. If we take long-range unscreened Coulomb potentials as the scattering potentials, the Fourier transform is proportional to q^{-1}, i.e., the inverse wave vector change during scattering, resulting in a E^{-2}-dependence of the matrix element on energy. This in turn leads to $\tau_0(E) \propto E$. Also, in the definition of the mobility, the effective mass has to be replaced by the relativistic mass $E/c^{\star 2}$ such that the energy dependence cancels, and the mobility becomes independent of density (Fermi energy), as observed in the experiment.

Although reasons for an energy-independent conductivity can be found, the details of charge transport and scattering mechanisms in graphene have remained a topic of research and discussion to date. Also, the question of why the minimum conductivity arises around $4e^2/h$ is not completely clear yet. It seems to emerge that as the Fermi energy comes close to the charge neutrality point, the electronic system consists of a random network of electron and hole puddles (Martin *et al.*, 2008), but strong localization cannot occur as a result of pseudorelativistic Klein tunneling (Katsnelson *et al.*, 2006) and the lack of a band gap.

Further reading

- Drude theory, Boltzmann equation: Shockley 1950; Seeger 2004; Balkanski and Wallis 2000.
- Papers: Drude 1900*a*; Drude 1900*b*; van der Pauw 1958*a*; van der Pauw 1958*b*.
- Scattering mechanisms: Ferry 1998; Davies 1998.

Exercises

(10.1) Consider a block of pure copper of dimension $L_x \times L_y \times L_z$.

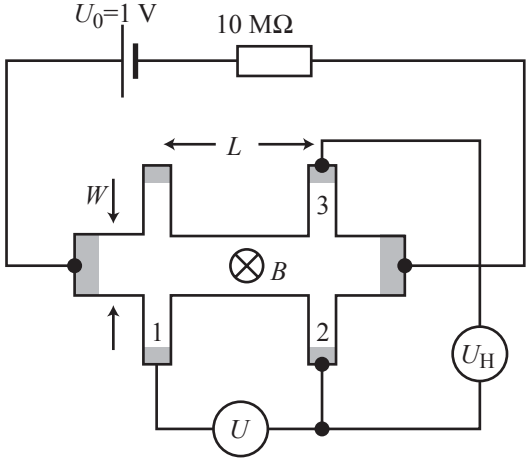
(a) Estimate the density of conduction band electrons in this three-dimensional system.

(b) A current flows in the x-direction and a magnetic field B is applied in the z-direction. What is the Hall resistance of the block?

(c) How small does L_z have to be in order to have a Hall resistance of the order of e^2/h?

(d) Discuss the characteristic differences between such a thin copper sheet and a two-dimensional electron gas.

(10.2) The measurement setup depicted below is used to measure the resistance of a two-dimensional electron gas in GaAs ($m^\star = 0.067m$) at the temperature $T = 4.2\,\mathrm{K}$. The Hall bar sample and a 10 MΩ resistor are connected in series to a voltage source that delivers the voltage $U_0 = 1\,\mathrm{V}$.



The total resistance of the sample is small compared to 10 MΩ. A magnetic field B can be applied normal to the plane of the two-dimensional electron gas. Between contacts 1 and 2 (separation $L = 100\,\mu\mathrm{m}$), the voltage $U = 10\,\mu\mathrm{V}$ is measured at zero magnetic field. Between contacts 2 and 3 the Hall voltage $U_\mathrm{H} = 200\,\mu\mathrm{V}$ is measured at $B = 1\,\mathrm{T}$. The width of the sample is $W = 30\,\mu\mathrm{m}$.

(a) What is the longitudinal (specific) resistivity ρ_{xx} and the transverse resistivity ρ_{xy} of the electron gas?

(b) Calculate the mobility μ, the scattering time τ, and the mean free path l of the electron gas at the Fermi energy from the measured resistivities.

(10.3) Show that the Einstein relation (10.54) holds for arbitrary dimensions and dispersions, if $\mathcal{D}_\mathrm{2D}$ is replaced by $\mathcal{D}(E_\mathrm{F})$. Show that, for parabolic dispersions, the diffusion constant can be expressed as $D = v_\mathrm{F}^2\tau/d$, if d is the dimensionality of the system.

(10.4) Discuss the two subband case starting from eq. (10.49). Set up a system of two equations for determining the scattering times $\tau_{1,2}^{\pm}$.

Ballistic electron transport in quantum point contacts

11

11.1 Experimental observation of conductance quantization

When we discussed the self-consistent calculation of the potential and the modes in an infinite wire (section 8.1), we saw that the number of occupied modes can be tuned with the voltage applied between the gate electrodes and the two-dimensional electron gas. Experimentally, short wires can be realized in split-gate structures (see Fig. 6.11, and the inset of Fig. 11.1) placed on top of a Ga[Al]As heterostructure incorporating a two-dimensional electron gas. If a negative voltage is applied to the gates, the electron gas below the gates can be depleted and a narrow channel remains connecting the two large two-dimensional electron reservoirs.

In 1988 two experiments by van Wees and co-workers, and Wharam and co-workers, on the low-temperature conductance of such quantum point contacts at zero magnetic field showed remarkable results. Figure 11.1(a) shows the measured resistance as a function of the voltage applied to the split gate. The measured resistance increases as the voltage is decreased, in agreement with the intuition that the width of the channel decreases. However, the resistance increase shows pronounced steps once the resistance value exceeds a few kΩ. Detailed investigations of this behavior showed that the two-dimensional electron gas connecting the quantum point contact to the external ohmic contacts contributes a gate-voltage independent series resistance of 400 Ω. If this resistance is subtracted, the resistance plateaus appear at quantized values $h/2Ne^2$, where N is an integer number. The conductance, determined as the inverse of the resistance is shown in Fig. 11.1(b). It shows pronounced plateau values at

$$G = \frac{2e^2}{h}N, \tag{11.1}$$

where N is an integer number. This result implies that the conductance is quantized in units of twice the conductance quantum

$$G_0 = \frac{e^2}{h} = 3.8740459 \times 10^{-5}\ \Omega^{-1}. \tag{11.2}$$

Since its discovery, the quantization of the conductance has been observed in a large number of experiments on samples of vastly differ-

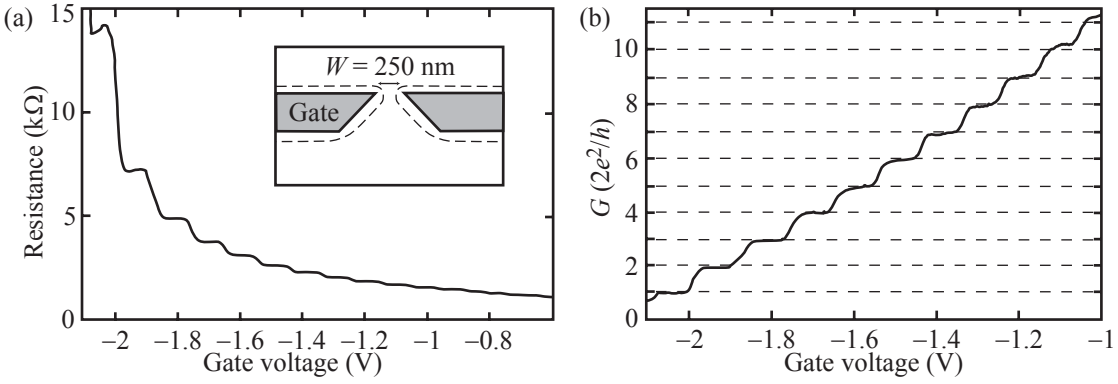


Fig. 11.1 (a) Resistance of a quantum point contact as a function of the gate voltage. The inset shows a schematic top view of the split-gate structure. (b) Conductance of the same quantum point contact as a function of gate voltage after subtraction of a gate voltage independent series resistance of 400 Ω. (Reprinted with permission from van Wees *et al.*, 1988. Copyright 1988 by the American Physical Society.)

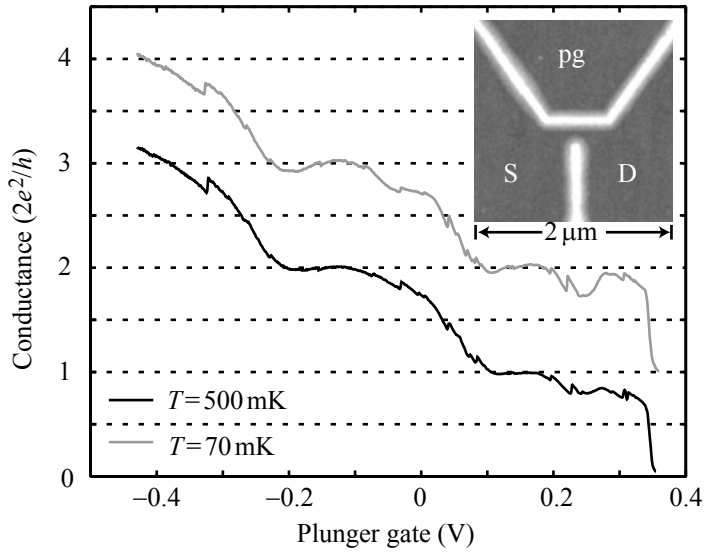


Fig. 11.2 Conductance quantization in a quantum point contact fabricated by AFM lithography on a *p*-type GaAs heterostructure (see inset). The labels 'pg', 'S', and 'D' denote the plunger gate, the source, and the drain contacts, respectively. The 70 mK curve is offset by $2e^2/h$ for clarity.

ent materials. Figure 11.2 shows an example of the effect observed on a quantum point contact fabricated in a two-dimensional hole gas in GaAs. Small kinks on the measured curve are most likely the result of rearrangements of charge in the sample close to the quantum point contact arising as the gate voltage is swept. The sample shown in the inset was fabricated by AFM lithography. The experimental conditions for the observability of the quantization are samples of high quality in which the electron (or hole) mean free path is very large compared to the length and width of the channel. In order to observe the quantization, the width of the channel must be comparable to the Fermi wavelength of the electrons, and the temperature must be low compared to the characteristic energy spacing of transverse modes in the channel.

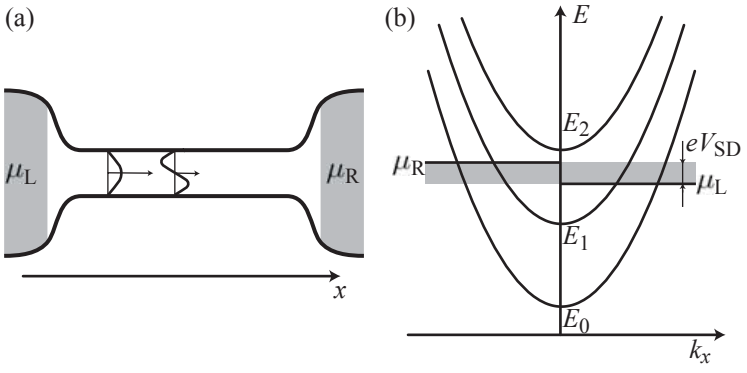


Fig. 11.3 (a) One-dimensional channel connected to left and right electron reservoirs (gray). The electrochemical potentials of the reservoirs are μ_L and μ_R. Transverse modes are schematically drawn within the channel with arrows indicating their propagation direction. (b) Dispersion relation in the one-dimensional channel. For each mode n, the parabolic dispersion relation has its minimum at energy E_n. States with negative k_x propagate from right to left and are fed from the right reservoir with electrochemical potential μ_R, while those with negative k_x travel from left to right and are fed from the left reservoir with electrochemical potential μ_L. The gray-shaded energy interval is given by the applied voltage between left and right reservoirs (bias window).

11.2 Current and conductance in an ideal quantum wire

In order to understand the experimental finding of conductance quantization, we consider the simple model of a perfect one-dimensional channel, as produced, for example, in the split-gate device in a Ga[Al]As heterostructure used for the experiments described above. Such a channel is schematically depicted in Fig. 11.3(a). In order to simplify the reasoning, we take the channel to be very long compared to its cross-sectional area, such that it can be treated in good approximation as being translationally invariant in the x-direction. Our goal is to find the current through this ideal wire in response to a voltage applied between two big electron reservoirs connecting to the wire. The quantum problem for the states in the wire is separable and we can write the wave functions in the wire as

$$\psi_{n\mathbf{k}}(\mathbf{r}) = \chi_n(y,z) \cdot \frac{1}{\sqrt{L}} e^{ik_x x}, \tag{11.3}$$

where L is a normalization length (very large compared to the relevant electronic wavelengths), and $\chi_n(y,z)$ are quantized states normal to the wire direction. These states are called the *transverse modes* of the wire. We assume a parabolic energy dispersion along the wire

$$E_n(k_x) = E_n + \frac{\hbar^2 k_x^2}{2m^\star},$$

where the E_n are contributions to the energies arising due to mode quantization normal to the wire axis. Therefore, the quantum number n labels the modes of the quantum wire. Positive values of k_x denote states

propagating from left to right, negative k_x those traveling from right to left. This dispersion relation is schematically shown in Fig. 11.3(b).

We now determine an expression for the contribution to the electrical current produced by an electron in a particular state (n, k_x). The quantum mechanical expression for the current density $\mathbf{j}_{nk_x}(\mathbf{r})$ is

$$\mathbf{j}_{nk_x}(\mathbf{r}) = -\frac{|e|\hbar}{2im^\star}\left(\psi^\star_{nk_x}(\mathbf{r})\nabla\psi_{nk_x}(\mathbf{r}) - \psi_{nk_x}(\mathbf{r})\nabla\psi^\star_{nk_x}(\mathbf{r})\right).$$

Inserting the wave function (11.3) leads to

$$d\mathbf{j}_{nk_x}(\mathbf{r}) = -\frac{|e|}{L}\left|\chi_n(y,z)\right|^2\frac{\hbar k_x}{m^\star}\mathbf{e}_x,$$

where $\mathbf{e}_x$ is a unit vector in the wire direction. The inverse normalization volume L is related to the k_x-interval $dk_x = 2\pi/L$ between two successive k_x-values leading to

$$d\mathbf{j}_{nk_x}(\mathbf{r}) = -\mathbf{e}_x\frac{|e|}{2\pi}\left|\chi_n(y,z)\right|^2\frac{\hbar k_x}{m^\star}dk_x. \tag{11.4}$$

The quantity $d\mathbf{j}_{nk_x}$ is therefore the infinitesimal contribution of a small interval dk_x to the current density. This equation is equivalent to the well-known expression $\mathbf{j} = \rho\mathbf{v}$ occurring in electrodynamics or fluid mechanics if we identify the charge density to be $\rho = -|e|dk_x\left|\chi_n(y,z)\right|^2/2\pi$ and the expectation value of the velocity in the x-direction

$$\mathbf{v}_n(k_x) = \mathbf{e}_x\frac{\hbar k_x}{m^\star} = \mathbf{e}_x\langle nk_x|\frac{\partial H}{\partial p_x}|nk_x\rangle = \mathbf{e}_x\frac{1}{\hbar}\frac{\partial E_n(k_x)}{\partial k_x}.$$

If we use the above expression for the velocity, we obtain, for the current density,

$$d\mathbf{j}_{nk_x}(\mathbf{r}) = -\mathbf{e}_x g_\mathrm{s}\frac{|e|}{h}\left|\chi_n(y,z)\right|^2\frac{\partial E_n(k_x)}{\partial k_x}dk_x.$$

Here we have introduced the spin degeneracy factor g_s, which takes the value $g_\mathrm{s} = 2$ in the case of GaAs. We convert the small interval dk_x into a small energy interval by using $dk_x = dE\partial k_x/\partial E_n(k_x)$, and realize that the expression $\partial k_x/\partial E_n(k_x)$ represents the one-dimensional density of states up to a factor 2π, but it is inversely proportional to the velocity $\partial E_n(k_x)/\partial k_x$. We therefore obtain

$$d\mathbf{j}_{nE}(\mathbf{r}) = \mp\mathbf{e}_x g_\mathrm{s}\frac{|e|}{h}\left|\chi_n(y,z)\right|^2 dE,$$

where the energy dependence of the density of states has exactly canceled the energy dependence of the group velocity. The minus sign is valid for right-moving states with $k_x > 0$, whereas the positive sign is applicable for $k_x < 0$. The exact cancelation of group velocity and one-dimensional density of states leads to the remarkable result that the contribution of a small energy interval dE to the current density is independent of the absolute value of the energy. Small group velocities at low energies are

exactly compensated by the large density of states at low energies, and large group velocities at high energies are exactly compensated by the lower density of states at these higher energies. This exact cancelation will turn out to be the key to the quantization of the conductance, as we will see below.

The current contributions $d\mathbf{I}_n(E)$ of states within a small energy interval are obtained from the previous result by integration over the cross-section of the wire. Using the fact that the transverse modes $\chi_n(y,z)$ are normalized, we obtain

$$d\mathbf{I}_n(E) = \mp \mathbf{e}_x g_\mathrm{s} \frac{|e|}{h} dE. \tag{11.5}$$

This contribution to the current is again independent of the energy around which states are considered, but also independent of the mode index n. The fundamental proportionality constant $|e|/h$ is the inverse of the magnetic flux quantum $\phi_0 = h/|e|$. Taking into account that $dE/|e|$ has the units of an electric voltage, we can identify e^2/h to be the fundamental quantum unit of electrical conductance G_0 in eq. (11.2).

Electrical current in thermodynamic equilibrium. Starting from eq. (11.5) we are now able to calculate the total current in the wire by energy integration. It is now important to notice that right-moving states (those with positive k_x) are occupied via the left electron reservoir connecting to the wire, whereas left-moving states (negative k_x) are occupied via the right electron reservoir [cf., Figs. 11.3(a) and (b)]. If the reservoirs are in thermodynamic equilibrium with each other, left- and right-moving states will both be occupied according to the same equilibrium Fermi–Dirac distribution function, and the total current is

$$I_\mathrm{tot}^{(\mathrm{eq})} = g_\mathrm{s} \frac{|e|}{h} \left(\sum_n \int_{E_n}^\infty dE f(E) - \sum_n \int_{E_n}^\infty dE f(E) \right) = 0.$$

The current contributions from left to right and vice versa cancel exactly, leading to zero total current.

Nonequilibrium currents due to an applied voltage. If the left and right electron reservoirs are not in thermodynamic equilibrium, for example because a voltage is applied between them via an external voltage source, the distribution functions for left- and right-moving electrons will differ and the net current is given by

$$I_\mathrm{tot} = g_\mathrm{s} \frac{|e|}{h} \sum_n \int_{E_n}^\infty dE \left[f_\mathrm{L}(E) - f_\mathrm{R}(E) \right], \tag{11.6}$$

where the subscripts L and R refer to the left and the right reservoirs. The distribution functions are given by the Fermi–Dirac distributions

$$f_i(E) = \frac{1}{\exp\left(\frac{E-\mu_\mathrm{i}}{k_\mathrm{B}T}\right) + 1}.$$

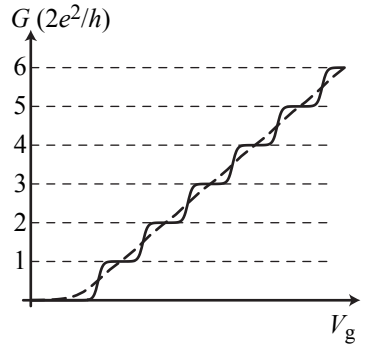


Fig. 11.4 Conductance as a function of gate voltage V_g according to eq. (11.8), assuming that the mode energies shift down proportional to the gate voltage. The spin degeneracy has been assumed to be $g_s = 2$. If the temperature is significantly lower than the spacing between the mode energies, the conductance is quantized (solid line). If $4k_BT$ is comparable to the mode's energy spacing, the quantization is smoothed out (dashed curve).

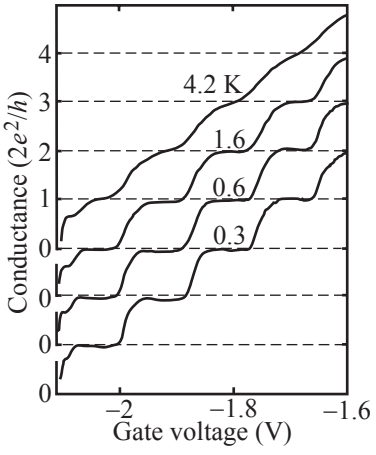


Fig. 11.5 Experimentally observed temperature dependence of the conductance quantization in a quantum point contact. (Reprinted with permission from van Wees *et al.*, 1991. Copyright 1991 by the American Physical Society.)

Linear response. Assuming that the applied voltage difference V_{SD} is small, i.e., $\mu_L - \mu_R = |e|V_{SD} \ll k_BT$, we can expand

$$f_L(E) - f_R(E) = \frac{\partial f_L(E)}{\partial \mu_L}(\mu_L - \mu_R) = -\frac{\partial f_L(E)}{\partial E}|e|V_{SD}. \qquad (11.7)$$

Inserting into eq. (11.6) and performing the energy integration gives

$$I_{tot} = g_s \frac{e^2}{h} \sum_n f_L(E_n) V_{SD}.$$

As a consequence, we obtain the quantization of the linear conductance of an ideal one-dimensional channel

$$G = \frac{I_{tot}}{V_{SD}} = g_s \frac{e^2}{h} \sum_n f_L(E_n) \qquad (11.8)$$

The behavior of this expression is shown in Fig. 11.4 for the case that the mode energies E_n show a linear dependence on the applied gate voltage. If the energy E_n of a particular mode n is well below the Fermi energy, the Fermi–Dirac distribution in the sum of eq. (11.8) is essentially one, the mode is occupied, and it contributes an amount $g_s e^2/h$ to the conductance. In turn, if the energy E_n is well above the Fermi energy, the Fermi–Dirac distribution is zero, the mode is not occupied, and it does not contribute to the conductance. Figure 11.4 resembles the most striking features of the experimental results in Fig. 11.1(b). If the separation between the mode energies E_n is significantly larger than k_BT, the conductance increases in steps of $2e^2/h$ with each additional occupied mode. The sharpness of the steps is in this ideal model given by k_BT. If the separation between mode energies E_n is comparable to, or smaller than $4k_BT$, the quantized conductance is completely smoothed out, in agreement with experimental observations. Figure 11.5 shows the experimental temperature dependence of the conductance of a quantum point contact. The temperature dependence of the quantization can be used to estimate the energetic separation of the lateral modes in the point contact.

In the limit of zero temperature, the conductance becomes

$$G = \frac{I_{tot}}{V_{SD}} = g_s \frac{e^2}{h} N,$$

where N is the number of occupied modes, in agreement with eq. (11.1). This important result implies that, in an ideal quantum wire with N occupied modes at zero temperature, each mode contributes one conductance quantum e^2/h to the total conductance. The number of occupied modes can be estimated by comparing the width of the channel and the Fermi wavelength of the electrons to be $N \approx 2W/\lambda_F$.

Resistance and energy dissipation. A remarkable property of the above result is that the conductance of the ideal channel is finite even

though there is no scattering inside. From Ohm's law, however, we are used to the fact that any finite resistance comes along with energy dissipation. We therefore have to answer the question of where energy is dissipated in our system, and how this energy dissipation leads exactly to the conductance quantum e^2/h. We answer these questions with the help of Fig. 11.6. In this figure we can see that an electron leaving the left contact from an energy below $\mu_{\rm L}$ leaves a nonequilibrium hole behind. In addition, the electron constitutes a nonequilibrium charge carrier in the drain at an energy above $\mu_{\rm R}$, after it has traversed the ideal quantum wire. The electron in the drain, and the hole in the source contact, will eventually relax to the respective electrochemical potential. The energy dissipated in the source contact is $\mu_{\rm L} - E$ and the energy dissipated in the drain contact amounts to $E - \mu_{\rm R}$. As a consequence, the total energy dissipated by a single electron having traversed the wire without scattering is $(\mu_{\rm L} - E) + (E - \mu_{\rm R}) = \mu_{\rm L} - \mu_{\rm R} = |e|V_{\rm SD}$. Most importantly, the dissipated energy for one electron is independent of the energy of this electron. In order to find the total power dissipated in the system due to all electrons traversing the wire at all energies between $\mu_{\rm R}$ and $\mu_{\rm L}$ we argue as follows: a single electron dissipates a power

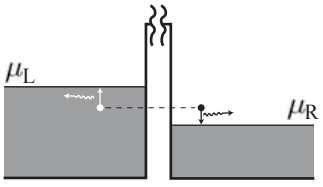


Fig. 11.6 Energy dissipated by a single electron that travels without scattering from source to drain contact through the wire. The wire itself, located between the two reservoirs, is omitted in the drawing for clarity.

$$dP = \frac{|e|V_{\rm SD}}{\tau},$$

where τ is the time that a single electron needs on average to traverse the wire. We obtain this time from the electrical current which counts the number of electrons traversing the wire per unit time. Using eq. (11.5) we have

$$dI_{\rm tot} = g_{\rm s}\frac{|e|}{h}NdE = \frac{|e|}{\tau} \Rightarrow \frac{1}{\tau} = \frac{g_{\rm s}NdE}{h},$$

giving the contribution of the infinitesimal energy interval dE to the dissipated power

$$dP = \frac{g_{\rm s}N|e|V_{\rm SD}}{h}dE.$$

We assume zero temperature for simplicity, and integrate this expression between $\mu_{\rm R}$ and $\mu_{\rm L}$. The resulting total dissipated power is

$$P = \frac{g_{\rm s}N|e|V_{\rm SD}}{h}(\mu_{\rm L} - \mu_{\rm R}) = \frac{g_{\rm s}N|e|V_{\rm SD}}{h}|e|V_{\rm SD} = \frac{V_{\rm SD}^2}{h/e^2 g_{\rm s}N}.$$

The resistance appearing from this power dissipation argument is

$$R_{\rm c} = \frac{h}{e^2}\frac{1}{g_{\rm s}N}, \tag{11.9}$$

which corresponds exactly to the quantized conductance that we have derived before. We conclude that the finite quantized conductance of an ideal wire without scattering arises due to the power dissipation of the charge carriers in the source and drain contacts. The resistance $R_{\rm c}$ is therefore often called the *contact resistance*.

In linear transport we can define a characteristic inelastic scattering length $l_{\rm i}$. Closer than $l_{\rm i}$ to the wire (or the quantum point contact),

there is no equilibrium distribution of charge carriers in the contacts, whereas at much larger distances the equilibrium distribution is fully restored.

11.3 Current and transmission: adiabatic approximation

In the previous discussion we have assumed a translationally invariant wire. Strictly speaking, this assumption is not compatible with connecting the wire to big reservoirs which breaks translational invariance. In the experiment, the length of the electron channel is even quite small, not orders of magnitude bigger than the width. A more realistic description of conductance quantization has to take these complications into account. One way to do this is to use the adiabatic approximation as discussed by Yacoby and Imry 1990. Within this approximation we assume that the transition from the macroscopic electron reservoirs [gray regions in Fig. 11.3(a)] into the wire is very smooth on the scale of the Fermi wavelength. The system is then described with the single-particle hamiltonian

$$H = -\frac{\hbar^2}{2m^\star}\Delta + V(x,y,z).$$

This hamiltonian is now split into the two parts

$$\begin{aligned} H_x &= -\frac{\hbar^2}{2m^\star}\frac{\partial^2}{\partial x^2} \\ H_{yz}(x) &= -\frac{\hbar^2}{2m^\star}\left(\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right) + V(x,y,z), \end{aligned}$$

where the variable x in $H_{yz}(x)$ is regarded as a parameter. We now assume that we have solved the eigenvalue problem for $H_{yz}(x)$ and we have found the eigenfunctions $\chi(y,z;x)$ obeying the equation

$$H_{yz}(x)\chi_n(y,z;x) = E_n(x)\chi_n(y,z;x)$$

for any particular value of x. We regard the quantum number n as denoting modes.

Now we can expand the wave function $\psi(x,y,z)$ of the original three-dimensional eigenvalue problem for the hamiltonian H at each point x in orthonormalized eigenfunctions $\chi_n(y,z;x)$:

$$\psi(x,y,z) = \sum_n \xi_n(x)\chi_n(y,z;x).$$

The expansion coefficients $\xi_n(x)$ will obey an equation that we obtain by inserting the expansion into the eigenvalue problem for H leading to

$$\sum_n (H_x + E_n(x))\xi_n(x)\chi_n(y,z;x) = E\sum_n \xi_n(x)\chi_n(y,z;x),$$

and projecting into the subspace of a particular mode m. This amounts to multiplying by $\chi^\star_m(y,z;x)$ and integrating over y and z. The result is

$$\sum_n \int dydz \;\; \chi^\star_m(y,z;x) H_x \chi_n(y,z;x)\xi_n(x) + E_m(x)\xi_m(x) = E\xi_m(x).$$

Calculating the matrix elements of H_x we obtain

$$-\frac{\hbar^2}{2m^\star}\frac{\partial^2\xi_m(x)}{\partial x^2} + E_m(x)\xi_m(x) + \sum_n K_{nm}(x)\xi_n(x) + \sum_n p_{nm}(x)\frac{\partial\xi_n(x)}{\partial x} = E\xi_m(x), \quad (11.10)$$

where

$$\begin{aligned} p_{nm}(x) &= -\frac{\hbar^2}{m^\star}\int \chi^\star_m(y,z;x)\frac{\partial\chi_n(y,z;x)}{\partial x}\, dydz \\ K_{nm}(x) &= -\frac{\hbar^2}{2m^\star}\int \chi^\star_m(y,z;x)\frac{\partial^2}{\partial x^2}\chi_n(y,z;x)\, dydz. \end{aligned}$$

So far no approximations have been involved and the problem has simply been rewritten.

The adiabatic approximation neglects terms containing $K_{nm}(x)$ and $p_{nm}(x)$. This approximation is justified if the wave function $\chi_n(y,z;x)$ changes smoothly with x. It leads to the simplified one-dimensional problem

$$-\frac{\hbar^2}{2m^\star}\frac{\partial^2\xi_m(x)}{\partial x^2} + E_m(x)\xi_m(x) = E\xi_m(x),$$

in which the electron experiences the effective potential

$$V^{\rm eff}_m(x) = E_m(x).$$

The effective potential depends on the mode index n, meaning that the potential barrier seen by an electron traversing the point contact depends on the transverse mode of its wave function. Figure 11.7 shows an example of a hard-wall confinement potential in the x-y-plane. Figure 11.8 shows the corresponding $V^{\rm eff}_0(x)$. The closer the electron approaches to $x = 0$ the narrower is the channel and the larger is $V^{\rm eff}_0(x)$. In the adiabatic approximation, the original quantum wire geometry reduces to a one-dimensional potential barrier problem.

We now aim at calculating the current in the adiabatic approximation. To this end, we define the zero of the potential such that for $x \to \pm\infty$ we have $V^{\rm eff}_m(x) \to 0$. Far away from the point contact we will then have the asymptotic solutions of the one-dimensional quantum problem

$$\xi^+_m(x) = \frac{1}{\sqrt{L}}\begin{cases} e^{ik_x x} + r_m e^{-ik_x x} & \text{for } x \to -\infty \\ t_m e^{ik_x x} & \text{for } x \to \infty \end{cases}$$

and

$$\xi^-_m(x) = \frac{1}{\sqrt{L}}\begin{cases} e^{-ik_x x} + r'_m e^{ik_x x} & \text{for } x \to \infty \\ t'_m e^{-ik_x x} & \text{for } x \to -\infty. \end{cases}$$

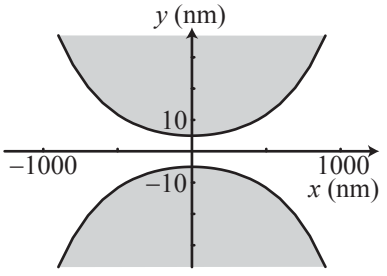


Fig. 11.7 Example of a hard-wall confinement potential in the x-y-plane forming a quantum point contact. Electron transport occurs in the x-direction; gray shaded areas are forbidden for electrons.

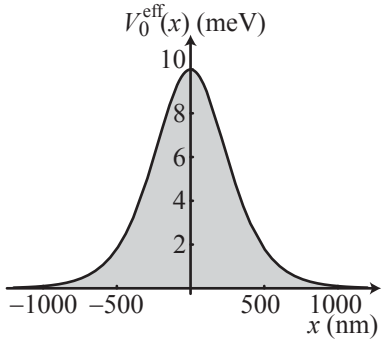


Fig. 11.8 Effective potential barrier $V^{\rm eff}_0(x)$ of the quantum point contact along x in the adiabatic approximation.

In complete analogy with eq. (11.4) we obtain the current density for right-moving and left-moving states

$$d\mathbf{j}^{\pm}_{mk_x}(y,z) = -\mathbf{e}_x \frac{|e|}{2\pi} \left|\chi_m(y,z;x\to\pm\infty)\right|^2 \frac{\hbar k_x}{m^\star} \left|t_m\right|^2 dk_x,$$

where the positive sign is valid for $k_x > 0$, and the negative sign for $k_x < 0$.

We continue in analogy to the discussion of the perfect one-dimensional wire and calculate the contribution of a small energy interval dE to the current. The drift velocity and the one-dimensional density of states cancel in the same way, and we find

$$d\mathbf{I}^{\pm}_m(E) = \mp \mathbf{e}_x g_\mathrm{s} \frac{|e|}{h} \left|t_m(E)\right|^2 dE,$$

in analogy with eq. (11.5). Treating a realistic device geometry in the adiabatic approximation leads to the appearance of the energy-dependent transmission probability $\mathcal{T}_m(E) = \left|t_m(E)\right|^2 \leq 1$ reducing the current contribution of a mode below the value given by eq. (11.5). Figure 11.9 shows the typical form of the transmission $\mathcal{T}_0(E)$ for the effective potential in Fig. 11.8.

Building on this result we calculate the total current through the constriction and find

$$I_{tot} = g_\mathrm{s} \frac{|e|}{h} \sum_n \int_{-\infty}^{+\infty} dE \;\; \mathcal{T}_n(E) \left[f_R(E) - f_L(E)\right]$$

in analogy with eq. (11.6), but now incorporating the finite energy-dependent transmission probability $\mathcal{T}_n(E)$. In the case of small applied source–drain voltage we obtain the linear response result for the conductance

$$G = g_s \frac{e^2}{h} \sum_n \int_{-\infty}^{+\infty} dE \;\; \mathcal{T}_n(E) \left(-\frac{\partial f_\mathrm{L}(E)}{\partial E}\right), \tag{11.11}$$

generalizing eq. (11.8). The energy derivative of the Fermi–Dirac distribution is sharply peaked around the Fermi energy at low temperature. If the transmission for a certain mode n at the Fermi energy is close to one, the mode will contribute with one full conductance quantum $g_\mathrm{s}e^2/h$. This will happen if the Fermi energy is more than a few $k_\mathrm{B}T$ above the maximum of the effective potential V_m^{eff}. However, if the Fermi energy is well below the maximum of V_m^{eff}, the transmission is close to zero and the mode does not contribute to the conductance. In this way, we recover the quantization of the conductance at low temperatures by treating the quantum point contact in the adiabatic approximation.

We note here that only equilibrium properties of the system close to the Fermi energy enter the expression for the linear conductance. The transmission is calculated from wave functions without taking the applied bias into account. The temperature dependence is given by the derivative of the equilibrium Fermi–Dirac distribution function. In this sense, the conductance is an equilibrium property of the system.

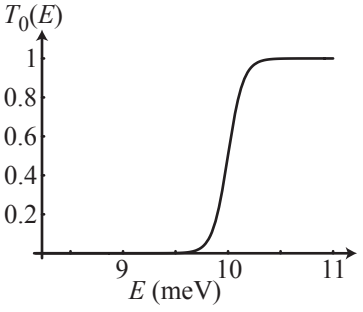


Fig. 11.9 Energy-dependent transmission $\mathcal{T}_0(E)$ for the energetically lowest mode.

At very low temperatures, the derivative of the Fermi–Dirac distribution is so sharp that the transmission does not change within a few $k_{\mathrm{B}}T$ around the Fermi energy. In this case, eq. (11.11) can be written as

$$G = g_s \frac{e^2}{h} \sum_n \mathcal{T}_n(E_F).$$

The smooth increase of the transmission $\mathcal{T}_n(E)$ for each mode n leads to smooth steps in the conductance even in the limit of zero temperature.

Power dissipation in the case of finite transmission. In the case of transmission one, for all occupied modes, the quantized conductance is recovered from the above considerations in the adiabatic case. The argument given for the power dissipation in the ideal wire case can be easily transferred to the adiabatic case considered here. The finite transmission of a particular mode n will reduce the rate at which an electron in this mode is transmitted by the factor $\mathcal{T}_n(E_{\mathrm{F}})$, or, in other words, it describes the delay in the transmission of the electron. The result is a reduction of the dissipated power compared to the ideal wire case, corresponding to the reduced conductance.

11.4 Saddle point model for the quantum point contact

The adiabatic approximation was general in the sense that no specific form of the potential $V(x, y, z)$ was assumed. As a consequence, the transmission coefficients (transmission amplitudes) t_m for the modes labeled with the letter m could not be worked out. The saddle point model of the quantum point contact assumes the potential to be of the form

$$V(x, y, z) = -\frac{1}{2} m^\star \omega_x^2 x^2 + \frac{1}{2} m^\star \omega_y^2 y^2 + V(z).$$

In this case the hamiltonian governing electron motion is separable. The solution for the motion in the z-direction is assumed to give quantized states with energies E_z having energy separations larger than any other energy scale in the problem. The confinement in the y-direction is parabolic and gives harmonic oscillator solutions with energies $E_y = \hbar\omega_y(m + 1/2)$. The total energy of an electron is $E = E_x + E_y + E_z$. The equation of motion for the x-direction is then

$$\left(-\frac{\hbar^2}{2m^\star} \partial_x^2 - \frac{1}{2} m^\star \omega_x^2 x^2 \right) \xi(x) = E_x \xi(x).$$

This equation can be rewritten as

$$\left(l_x^2 \partial_x^2 + \frac{x^2}{l_x^2} + \epsilon \right) \xi(x) = 0.$$

with the normalized energy scale $\epsilon = 2E_x/\hbar\omega_x$, and the length scale $l_x^2 = \hbar/m^\star\omega_x$. Solutions of this equation exist for arbitrary values of

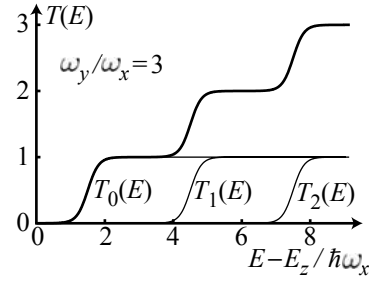


Fig. 11.10 Transmission of individual modes and total transmission in the saddle point model for the quantum point contact. (Reprinted with permission from Buttiker, 1990. Copyright 1990 by the American Physical Society.)

E_x. They can be written as linear combinations of parabolic cylinder functions $D_\nu(x)$, i.e.,

$$\xi(x) = c_1 D_{-\frac{1}{2}i(\epsilon-i)}[(1+i)x/l_x] + c_2 D_{\frac{1}{2}i(\epsilon+i)}[(-1+i)x/l_x].$$

The coefficients c_1 and c_2 can now be chosen in such a way that for $x \gg 0$ there is only a right-moving (transmitted) current, whereas for $x \ll 0$ there is an incoming and a reflected current. It has been shown (Miller, 1968; Fertig and Halperin, 1987; Buttiker, 1990) that the transmission of mode m through the parabolic potential barrier in the x-direction is then given by

$$\mathcal{T}_m(E) = \frac{1}{1+e^{-2\pi\epsilon_m}}, \tag{11.12}$$

with the energy parameter

$$\epsilon_m = \frac{E - \hbar\omega_y(m+1/2) - E_z}{\hbar\omega_x}.$$

Figure 11.10 shows the transmission of the lowest three modes and the total transmission. For energies $E \ll \hbar\omega_y(m+1/2)+E_z$, the transmission is exponentially suppressed. For the energy value $E = \hbar\omega_y(m+1/2)+E_z$, the transmission has the value of $1/2$, whereas it tends towards unity for $E \gg \hbar\omega_y(m+1/2)+E_z$. Well-defined plateaus form if the ratio ω_y/ω_x is well above one, because the width of the transition region between plateaus is set by the energy scale $\hbar\omega_x$, whereas the energy shift between neighboring plateaus is determined by $\hbar\omega_y$. If the conductance is calculated according to eq. (11.11), the temperature starts to dominate the width of the transitions between plateaus as soon as $k_\mathrm{B}T$ becomes larger than $\hbar\omega_x$. The quantization is severely smeared out at temperatures $k_\mathrm{B}T \gg \hbar\omega_y$.

The saddle point model for the quantum point contact can also be solved analytically if a magnetic field in the z-direction is present (Fertig and Halperin, 1987; Buttiker, 1990). Equation (11.12) remains valid, but the expression for ϵ_m has to be modified in order to account for the cyclotron energy scale.

11.5 Conductance in the nonadiabatic case

How well does the adiabatic approximation work if it is compared to an exact calculation? In order to get insight into this question, we look at a model proposed by Ulreich and Zwerger, 1998. Within this model, the point contact is described by a harmonic confinement potential. The confinement energy $\hbar\omega_y$ is varied along the x-direction from a small value through a maximum value at the point of the constriction and back down to the same small value leading to an effective width $b(x)$ of the quantum point contact. Figure 11.11 shows the geometry and the local density of states (i.e., the squared wave function modulus) at the Fermi energy for the adiabatic case (right) and the exact calculation (left). The depicted situation corresponds to the transition between the closed point

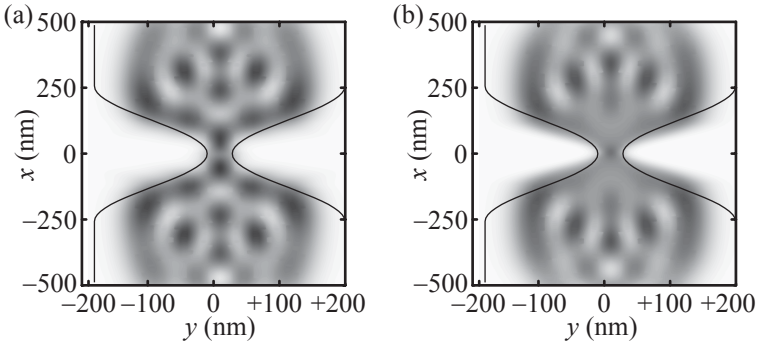


Fig. 11.11 (a) Local density of states in a quantum point contact with a harmonic confinement potential. The width of the confinement $b(x)$ varies with x. Close to the constriction, the adiabatic approximation (b) differs appreciably from the exact solution (a) (Ulreich and Zwerger, 1998).

contact and the first plateau at $2e^2/h$, where the point contact has a conductance of e^2/h, i.e., half the plateau value. Far away from the constriction, in the wide parts of the contacts, four modes exist. Very close to the constriction, the adiabatic solution shows much less fine structure than the exact solution that takes mixing between adiabatic modes into account. This mode mixing is caused by off-diagonal matrix elements $p_{nm}(x)$ and $K_{nm}(x)$ in eq. (11.10). In higher order approximations these matrix elements can be taken into account perturbatively (Yacoby and Imry, 1990).

Mode mixing means that there is a certain probability amplitude $t_{nm}(k_m)$ that an electron impinging on the quantum point contact in state (m, k_m) will be transmitted into state (n, k_n). As a result we have the asymptotic expressions for the wave functions

$$\psi_m^+(\mathbf{r}) = \frac{1}{\sqrt{L}} \begin{cases} \chi_m(y,z)e^{ik_m x} + \sum_n r_{nm}\chi_n(y,z)e^{-ik_n x} & \text{for } x \to -\infty \\ \sum_n t_{nm}\chi_n(y,z)e^{ik_n x} & \text{for } x \to +\infty \end{cases}$$

and

$$\psi_m^-(x) = \frac{1}{\sqrt{L}} \begin{cases} \chi_m(y,z)e^{-ik_m x} + \sum_n r'_{nm}\chi_n(y,z)e^{ik_n x} & \text{for } x \to +\infty \\ \sum_n t'_{nm}\chi_n(y,z)e^{-ik_n x} & \text{for } x \to -\infty \end{cases} .$$

Following again the previous scheme for calculating the right- and left-moving currents we find the total linear conductance

$$G = g_s \frac{e^2}{h} \int_0^\infty dE \left(-\frac{\partial f_L(E)}{\partial E} \right) \sum_{n,m} \mathcal{T}_{nm}(E), \tag{11.13}$$

where the transmission probabilities $\mathcal{T}_{nm}(E)$ from mode m into mode n at energy E are given by

$$\mathcal{T}_{nm}(E) := \frac{k_n(E)}{k_m(E)} \left| t_{nm}(E) \right|^2 .$$

Equation (11.13) can be seen as a general formula for the conductance of a two-terminal structure with noninteracting electrons, no matter whether it is a quantum point contact or another structure. The only ingredient for this expression is the asymptotic form of the wave function

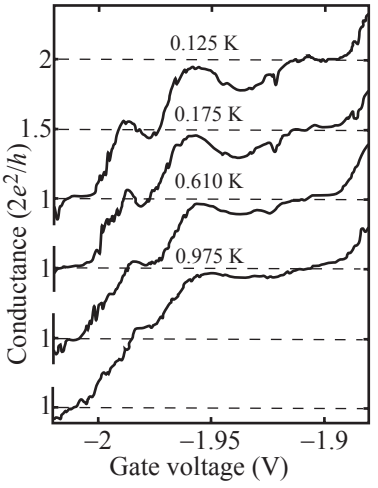


Fig. 11.12 Conductance of a quantum point contact showing transmission resonances. (Reprinted with permission from van Wees *et al.*, 1991. Copyright 1991 by the American Physical Society.)

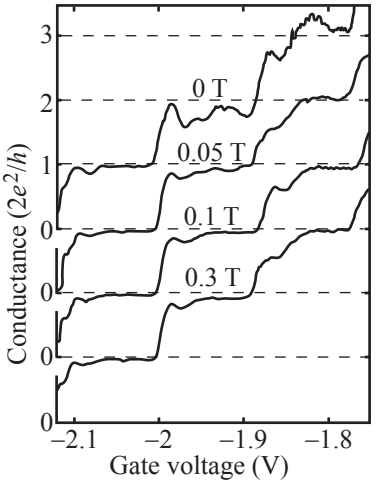


Fig. 11.13 Conductance of a quantum point contact showing the suppression of transmission resonances by a magnetic field. (Reprinted with permission from van Wees *et al.*, 1991. Copyright 1991 by the American Physical Society.)

in the leads. All the details of the potential landscape between the leads enter into the transmission probabilities $\mathcal{T}_{nm}(E)$. In a quantum point contact, strong scattering between different modes is detrimental to the observation of the conductance quantization.

11.6 Nonideal quantum point contact conductance

The conductance of a quantum point contact measured as a function of plunger gate voltage does not always show the perfect and smooth behavior shown in Fig. 11.1(b). For example, the changing gate voltage can lead to changes in the charge of impurity sites close to the constriction which leads to a change in the constriction potential, and thereby to a change in the transmission (see, for example, the kinks in the curves of Fig. 11.2).

Other nonideal behavior of the conductance can arise if the transmission of the constriction does not increase monotonically, but shows a modulation with energy. An example is shown in Fig. 11.12, where the transmission is modulated at the transition of the conductance from the first plateau at $2e^2/h$ to the second plateau at $4e^2/h$ in the curves taken at the lowest three temperatures. Such transmission modulation can occur if the microscopic potential does not form a smooth saddle shape, but has local potential minima with bound states through which electrons can be transmitted resonantly. Another possible reason for such a modulation of the transmission is coherent backscattering in the vicinity of the quantum point contact. Nonadiabatic coupling into and out of the quantum point contact can lead to imperfect quantization of the conductance visible on the plateaus.

Peculiarities in the conductance steps of quantum point contacts are frequently most pronounced at the lowest temperatures. In an intermediate temperature range, they tend to be smoothed by energy averaging, as shown in Fig. 11.12, where a decent plateau is visible at 0.975 K.

The modulation of the transmission can often be suppressed by a magnetic field normal to the plane of the electron gas, as can be seen in the experimental data in Fig. 11.13. If we assume that coherent backscattering is the origin of the modulated transmission, we can find a characteristic area from the magnetic field scale B_c on which the modulation is suppressed (this is related to the Aharonov–Bohm phase discussed later in this book). The result is $A = \phi_0/B_c \sim 41.4\ \mathrm{Tnm^2}/0.1\ \mathrm{T} = 414\ \mathrm{nm^2}$. The characteristic length scale $\sqrt{A} \approx 20$ nm is of the order of half the Fermi wavelength and strong localization effects may therefore play a role in the vicinity of the point contact.

11.7 Self-consistent interaction effects

So far we have limited our considerations to relating the problem of calculating the conductance to a transmission problem for charge carriers. Now we will discuss qualitatively how the electrostatic potential landscape is changed in response to the current flowing through the structure.

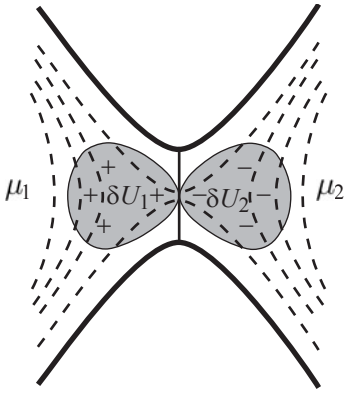


Fig. 11.14 Charge dipole developing at a quantum point contact under the influence of current flow. (Reprinted with permission from Christen and Buttiker, 1996. Copyright 1996 by the American Physical Society.)

Self-consistent screening and Landauer's resistance dipole. As a starting point we consider the many-body problem for the electronic system close to the quantum point contact to be solved. This can, for example, be achieved by using a self-consistent approach including Hartree and exchange energies, giving single-particle wave functions $\psi_n(\mathbf{r})$. In this situation, at zero magnetic field, no equilibrium currents will flow. The current flow through the system can now be regarded as a perturbation of this equilibrium situation to be described in first order perturbation theory. Under the influence of the current flow, the self-consistent potential of the system will slightly change. This is essentially due to the fact that charge carriers injected from the source contact and transmitted to the drain constitute a nonequilibrium contribution to the carrier density in the drain contact. At the same time these carriers are missing in the source contact, which also leads to a slight variation from thermodynamic equilibrium. The result of these self-consistently adjusting nonequilibrium carrier densities is a charge dipole schematically depicted in Fig. 11.14, and called *Landauer's resistivity dipole*.

The screened electrostatic potential of this dipole arises as an additional potential around the quantum point contact. However, the far field of the dipole potential tends to zero with increasing distance from the point contact and no electrostatic voltage drop results across the point contact.

11.8 Diffusive limit: recovering the Drude conductivity

The question arises as to how far, and under which approximations, the Landauer–Büttiker expression for the two-terminal conductance, given in eq. (11.13), is equivalent to the results of the Drude model. In order to clarify this point, we consider a wire with M occupied modes. The number of modes is related to the Fermi wavelength λ_F of the two-dimensional electron gas and the width W of the point contact constriction via $M = 2W/\lambda_F$. We further consider only a single scatterer in the sample with mean transmission $\mathcal{T}_1$ per mode at the Fermi energy. The zero-temperature conductance is then given by

$$G = \frac{2e^2}{h} M \mathcal{T}_1$$

and the resistance of the wire is

$$R = \frac{h}{2e^2}\frac{1}{M\mathcal{T}_1}.$$

Using the concept of the contact resistance, eq. (11.9), we can split the total resistance into

$$R = R_\mathrm{c} + \frac{h}{2e^2}\frac{1}{M}\frac{1-\mathcal{T}_1}{\mathcal{T}_1} = R_\mathrm{c} + R_\mathrm{W}.$$

While the contact resistance arises due to inelastic relaxation in the contacts in the case of the ideal wire, the second term represents the pure resistance contribution of the wire itself which we denote by R_W. This resistance tends to zero for $\mathcal{T}_1 \to 1$ as expected intuitively for an ideal pure wire; conversely, the wire resistance tends to infinity for $\mathcal{T}_1 \to 0$.

We refine this interpretation by extending our considerations to a wire with two scatterers having average transmission probabilities $\mathcal{T}_1$ and $\mathcal{T}_2$ with the corresponding reflection probabilities $\mathcal{R}_1$ and $\mathcal{R}_2$. Assuming that phase-coherence gets lost on a length scale smaller than the separation of the two scatterers, we can write for the total transmission

$$\begin{aligned}
\mathcal{T} &= \mathcal{T}_1\mathcal{T}_2 + \mathcal{T}_1\mathcal{R}_2\mathcal{R}_1\mathcal{T}_2 + \mathcal{T}_1\mathcal{R}_2\mathcal{R}_1\mathcal{R}_2\mathcal{R}_1\mathcal{T}_2 + \ldots \\
&= \mathcal{T}_1\mathcal{T}_2\sum_{n=0}^{\infty}(\mathcal{R}_1\mathcal{R}_2)^n = \frac{\mathcal{T}_1\mathcal{T}_2}{1-\mathcal{R}_1\mathcal{R}_2} \\
&= \frac{\mathcal{T}_1\mathcal{T}_2}{1-(1-\mathcal{T}_1)(1-\mathcal{T}_2)} = \frac{\mathcal{T}_1\mathcal{T}_2}{\mathcal{T}_1+\mathcal{T}_2-\mathcal{T}_1\mathcal{T}_2}.
\end{aligned}$$

From this result we obtain

$$\begin{aligned}
\frac{1-\mathcal{T}}{\mathcal{T}} &= \frac{\mathcal{T}_1+\mathcal{T}_2-2\mathcal{T}_1\mathcal{T}_2}{\mathcal{T}_1\mathcal{T}_2} \\
&= \frac{1-\mathcal{T}_1}{\mathcal{T}_1} + \frac{1-\mathcal{T}_2}{\mathcal{T}_2}.
\end{aligned}$$

The quantity $(1-\mathcal{T}_i)/\mathcal{T}_i$ is additive for an *incoherent* sequence of several scatterers in series. For an incoherent addition of N pieces of wire with one scatterer in each we obtain

$$\frac{1-\mathcal{T}(N)}{\mathcal{T}(N)} = N\left\langle\frac{1-\mathcal{T}_i}{\mathcal{T}_i}\right\rangle$$

with $\langle\ldots\rangle$ denoting the average over scatterers, and therefore

$$\mathcal{T}(N) = \frac{1}{N\langle(1-\mathcal{T}_i)/\mathcal{T}_i\rangle + 1}.$$

If we express the number of scatterers as $N = \nu L$, where ν is the density of scatterers per unit length, and L is the length of the wire, then we can write

$$\mathcal{T}(N) = \frac{\langle l_\mathrm{e}\rangle}{L + \langle l_\mathrm{e}\rangle} \tag{11.14}$$

with the new length scale $\langle l_\mathrm{e}\rangle = (\nu\langle(1-\mathcal{T}_i)/\mathcal{T}_i\rangle)^{-1}$. The meaning of this length scale becomes evident if we assume that $\mathcal{R}_i = 1-\mathcal{T}_i \ll 1$ and realize that then $\langle l_\mathrm{e}\rangle \approx 1/\langle\mathcal{R}_i\rangle\nu$. The quantity $\langle\mathcal{R}_i\rangle\nu$ can be seen as the average backscattering probability per unit length. The probability p_bs that an electron has not been backscattered after travelling a distance L therefore obeys the equation

$$\frac{dp_\mathrm{bs}(L)}{dL} = -\langle\mathcal{R}_i\rangle\nu p_\mathrm{bs}(L)$$

with the exponentially decaying solution

$$p_\mathrm{bs}(L) = e^{-\langle\mathcal{R}_i\rangle\nu L} = e^{-L/\langle l_\mathrm{e}\rangle}.$$

As we have suggested already by the choice of notation, the quantity $\langle l_\mathrm{e}\rangle$ can therefore be seen as a one-dimensional mean free path of the electron which should be of the order of the elastic mean free path l_e in Drude's theory. With this result, the total resistance of N sections of wire becomes

$$R_\mathrm{tot} = \frac{h}{2e^2}\frac{1}{M}\left(1+\frac{1-\mathcal{T}(N)}{\mathcal{T}(N)}\right).$$

Above, we have denoted the second term as being the pure resistance of the wire, R_W. We now take a closer look at this contribution and replace $\mathcal{T}(N)$ by the expression in eq. (11.14) containing the mean free path $\langle l_\mathrm{e}\rangle$. We find $\mathcal{T}(N)/(1-\mathcal{T}(N)) = L/\langle l_\mathrm{e}\rangle$. We further replace M by $2W/\lambda_\mathrm{F} = k_\mathrm{F}W/\pi$ (two dimensions) and obtain

$$R_\mathrm{W} = \frac{L}{W}\frac{h}{2e^2}\frac{1}{k_\mathrm{F}}\frac{\pi}{\langle l_\mathrm{e}\rangle}.$$

The pure resistance of the wire appears to be proportional to the length of the wire and inversely proportional to its width, as we know it for ohmic resistors. Identifying the elastic mean free path $l_\mathrm{e} = \langle l_\mathrm{e}\rangle/\pi$, the specific resistivity of the wire is

$$\rho_\mathrm{W} = \frac{h}{e^2}\frac{1}{k_\mathrm{F}l_\mathrm{e}}$$

and the specific conductivity is

$$\sigma_\mathrm{W} = \frac{e^2}{h}k_\mathrm{F}l_\mathrm{e}.$$

These expressions are identical to those known from the Drude model for the conductivity. The mean free path that we introduced above has consequences that agree with intuition: if ν is very low, the mean free path is very large. If the average reflection $\langle\mathcal{R}_i\rangle$ of a scatterer is very small, l_e will be very large. The contact resistance R_c will be negligible compared to R_W, if the wire length is large.

In summary, we can state that the resistance of a one-dimensional wire with M modes can be regarded as the series connection of the contact resistance R_c and the pure wire resistance R_W. The latter is

proportional to the length and inversely proportional to the width of the wire. The specific resistivity of the wire following from the Landauer–Büttiker description is in agreement with the Drude description, if there is a sufficient number of scatterers in the wire (diffusive limit) and if individual scattering segments are combined incoherently.

Further reading

- Landauer–Büttiker formalism: Datta 1997; Beenakker and van Houten 1991; Imry 2002.
- Papers: van Wees *et al.* 1988; Wharam *et al.* 1988; Landauer 1989; Payne 1989.

Exercises

(11.1) Calculate the transmission $\mathcal{T}(E)$ of a rectangular barrier potential (height V_0, width W) as a function of the energy of the incident electron. Discuss $\mathcal{T}(E)$ for $E < V_0$ and $E > V_0$. Why is $\mathcal{T}(E)$ not one for all $E > V_0$? What can you learn from this result for the strongly nonadiabatic transmission through quantum point contacts?

(11.2) Consider the quantum states in a planar quantum system of variable width $W(x) = ax^2/(x^2+1) + b$ $(a, b > 0)$. Calculate the effective one-dimensional barrier potential $V_n^{\mathrm{eff}}(x)$ that an electron in the lateral mode n (n is the number of nodes in the y-direction) experiences in the adiabatic approximation. Sketch this potential for $n = 0, 1, 2$.

(11.3) Within the Drude model, the conductance G_{D} of a diffusive wire made from a two-dimensional electron gas in the quantum limit is given by

$$G_{\mathrm{D}} = \frac{W}{L}\frac{n_{\mathrm{s}}e^2\tau}{m^\star},$$

where W is the width of the wire, L is its length, n_{s} is the sheet electron density, and τ is the Drude scattering time. Within the framework of the Landauer–Büttiker theory, the corresponding conductance can be written as

$$G_{\mathrm{LB}} = \frac{e^2}{h}M\mathcal{T},$$

where M is the number of occupied modes in the wire, and $\mathcal{T}$ is the average transmission per mode.

(a) Estimate the number of modes in the wire, given the width W and the Fermi wavelength λ_{F}.

(b) Estimate the average transmission $\mathcal{T}$ per mode by comparing G_{D} and G_{LB}. Hint: rewrite G_{D} in terms of the mean free path l and the Fermi wavelength λ_{F}.

(11.4) Consider a perfectly ballistic very long quantum wire connected to two big electron reservoirs. The central piece of the wire is a bit narrower, hosting only M modes, while the main part of the wire has $N > M$ modes occupied. Find reasons why we can attribute the (contact) resistance

$$R_{\mathrm{c}} = \frac{h}{2e^2}\left(\frac{1}{M} - \frac{1}{N}\right)$$

to the central piece of the wire. Hint: consider the situation assuming the adiabatic approximation to be valid. See also Landauer 1989 and Payne 1989 for further discussions.

Tunneling transport through potential barriers

12

In chapter 11 we introduced the concept of viewing the conductance of a nanostructure as being intimately related to the transmission of charge carriers. The concept was introduced in a very general way and, using only the saddle point model, one specific example was given, where the transmission can actually be calculated. Here we will look at ways of calculating the transmission through structures, where quantum tunneling is relevant. To get a feeling, we start with an example which is not of great practical interest, but it can be solved analytically. In the second step, we will show how tunneling can be treated using perturbative methods.

In semiconductor nanostructures, tunneling can occur under different circumstances. If a tunneling barrier is created in the conduction band of one particular direct III-V material, for example, in a split-gate device, and the barrier height is small compared to the separation to other bands, or band extrema of the material, only conduction band states at Γ will contribute, and the carriers can be treated as particles with the effective conduction band mass $m^\star$. This is the case we primarily have in mind during the calculations in this chapter. Complications arise, if electrons tunnel from material A through material B with a bigger bandgap into material A. For example, tunneling from GaAs through an AlAs barrier can involve the states at the X-minimum in AlAs, because it is lower in energy than the Γ-minimum. Yet another description is necessary if electrons tunnel from the valence band to the conduction band (interband tunneling). Again different situations can arise if electrons tunnel from a metal into a semiconductor at a Schottky contact. The general perturbative treatment of tunneling used below can also be applied to the calculation of the tunneling current between the metallic tip of a scanning tunneling microscope and a conducting surface.

12.1 Tunneling through a single delta-barrier

One of the simplest problems which captures the essence of quantum tunneling is scattering at a single delta-potential in one dimension. We consider the potential $U(x) = u\delta(x)$ describing a barrier at the origin (see Fig. 12.1). The wave function to the left of the barrier can then be

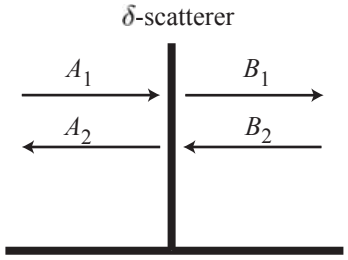


Fig. 12.1 Single delta-potential barrier scattering electrons waves. The quantities A_1, A_2, B_1, and B_2 are the amplitudes of the indicated incident and reflected plane wave states.

written as

$$\psi_{\rm L}(x) = A_1 e^{ikx} + A_2 e^{-ikx},$$

where $k = \sqrt{2mE/\hbar^2}$. To the right of the barrier we have correspondingly

$$\psi_{\rm R}(x) = B_1 e^{ikx} + B_2 e^{-ikx}.$$

The relation between the amplitudes A_i and B_i is found from the matching conditions at the position of the barrier

$$\psi(0^+) = \psi(0^-)$$

$$\psi'(0^+) - \psi'(0^-) = \frac{2mu}{\hbar^2}\psi(0).$$

Inserting the above wave functions leads to the transfer matrix T_k of the barrier which describes the relation between the incoming and outgoing amplitudes to the left of the scatterer with those right of the scatterer. It is found to be

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \underbrace{\frac{1}{2k}\begin{pmatrix} 2k - i\gamma & -i\gamma \\ i\gamma & 2k + i\gamma \end{pmatrix}}_{T_k} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix},$$

where $\gamma = 2mu/\hbar^2$. With $\alpha = 1 - i\gamma/2k$ and $\beta = i\gamma/2k$ we can write the transfer matrix as

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} \alpha & \beta^\star \\ \beta & \alpha^\star \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}. \tag{12.1}$$

The transmission amplitude t and the reflection amplitude r of the barrier can be obtained by letting $A_1 = 1$, $A_2 = r$, $B_1 = t$, and $B_2 = 0$. This gives

$$\begin{aligned} t &= \alpha + \beta^\star r \\ 0 &= \beta + \alpha^\star r \end{aligned}$$

and therefore

$$\begin{aligned} r &= -\frac{\beta}{\alpha^\star} = -\frac{i\gamma/2k}{1 + i\gamma/2k} \\ t &= \frac{|\alpha|^2 - |\beta|^2}{\alpha^\star} = \frac{1}{1 + i\gamma/2k} = \frac{1}{\alpha^\star}. \end{aligned}$$

The transmission probability $\mathcal{T} = |t|^2$ shown in Fig. 12.2 depends on the wave vector k and therefore on the energy of the incident particle. The expression

$$\mathcal{T} = \frac{1}{1 + \gamma^2/4k^2}$$

tends to zero for $k \to 0$ and to one for $k \to \infty$. At the characteristic value $k_{\rm c} = \gamma/2$, $\mathcal{T} = 1/2$, i.e., for $k \gg k_{\rm c}$ the barrier is transparent, whereas for $k \ll k_{\rm c}$ it is opaque.

The transmission probability $\mathcal{T}$ and the reflection probability $\mathcal{R} = |r|^2$ obey the well-known relation

$$\mathcal{T} + \mathcal{R} = |t|^2 + |r|^2 = \frac{1}{1 + (\gamma/2k)^2} + \frac{(\gamma/2k)^2}{1 + (\gamma/2k)^2} = 1.$$

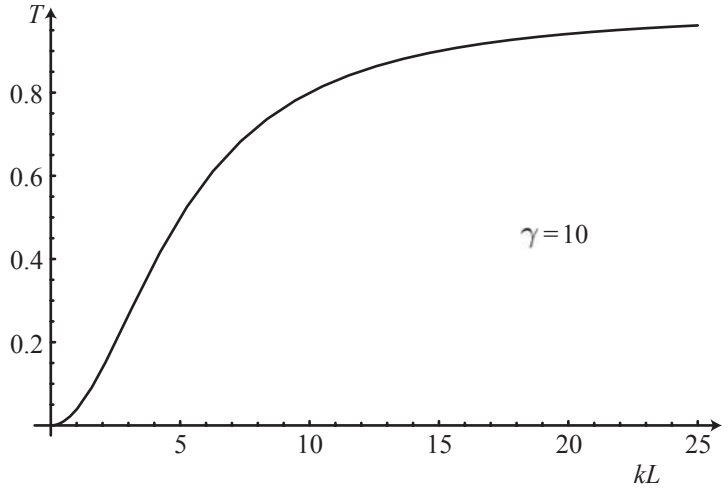


Fig. 12.2 Transmission probability $\mathcal{T}$ through a single delta-barrier as a function of the wave vector k for $\gamma = 10$.

12.2 Perturbative treatment of the tunneling coupling

In the following, we treat the case of weak tunneling coupling that can be treated perturbatively. Experimental realizations of such weakly coupled junctions are, for example, the quantum point contact that is almost pinched off, or tunneling junctions between the metallic tip of a scanning tunneling microscope and the conducting substrate.

The perturbative treatment of the tunneling coupling goes back to a paper of Bardeen (Bardeen, 1961), who introduced it for describing tunneling between metals (superconducting or normal) separated by a thin oxide tunneling barrier. It is, however, a very general way to describe tunneling between two materials. Although Bardeen's original theory has been made for interacting electronic systems, we will neglect interactions in the following for simplicity.

Assume that the potential $V(\mathbf{r}) < 0$ describes the tunneling barrier between two electronic systems. We introduce an (arbitrary) surface $\mathcal{S}$ cutting the barrier normal to the direction of current flow, which separates the problem into two spatially separate regions that we call left and right. Definition of this surface implies the definition of two characteristic functions

$$\theta_{\mathrm{L}}(\mathbf{r}) = \begin{cases} 1 & \text{for } \mathbf{r} \text{ left} \\ 0 & \text{elsewhere} \end{cases}$$

and

$$\theta_{\mathrm{R}}(\mathbf{r}) = \begin{cases} 1 & \text{for } \mathbf{r} \text{ right} \\ 0 & \text{elsewhere} \end{cases},$$

such that $\theta_{\mathrm{L}}(\mathbf{r}) + \theta_{\mathrm{R}}(\mathbf{r}) = 1$. In this way we split the potential into two contributions

$$V_{\mathrm{L}}(\mathbf{r}) = V(\mathbf{r})\theta_{\mathrm{L}}(\mathbf{r}) \quad \text{and} \quad V_{\mathrm{R}}(\mathbf{r}) = V(\mathbf{r})\theta_{\mathrm{R}}(\mathbf{r}),$$

with $V_{\rm R}(\mathbf{r}) + V_{\rm L}(\mathbf{r}) = V(\mathbf{r})$.

We write the hamilton operators for the two subsystems as

$$H_{\rm L} = T + V_{\rm L}(\mathbf{r}), \quad \text{and} \quad H_{\rm R} = T + V_{\rm R}(\mathbf{r}),$$

with T being the kinetic energies of the electrons. The eigenvalue problems of the two hamiltonians,

$$H_{\rm L}\phi_\mu(\mathbf{r}) = E_\mu\phi_\mu(\mathbf{r}) \quad \text{and} \quad H_{\rm R}\varphi_\nu(\mathbf{r}) = \epsilon_\nu\varphi_\nu(\mathbf{r}),$$

can now be solved separately. The wave functions $\phi_\mu(\mathbf{r})$ form a complete orthonormal basis in space (i.e., not only in the left region!). States with energies $E_\mu < 0$ will typically decay exponentially within the barrier, whereas states with $E_0 > 0$ are oscillatory functions also in the right region. Analogue statements can be made about the φ_ν. Pairs of wave functions φ_ν and ϕ_μ are not orthogonal to each other, but they possess a finite spatial overlap.

We are now interested in the electron dynamics described by the total hamilton operator

$$H = T + V(\mathbf{r}) = T + V_{\rm L}(\mathbf{r}) + V_{\rm R}(\mathbf{r}).$$

We assume that initially an electron occupies a particular state $\phi_0(\mathbf{r})$ in the left region with an energy $E_0 < 0$ at time $t = 0$. At later times $t > 0$ the electronic state will evolve into another state $\psi(\mathbf{r}, t)$. We expand $\psi(\mathbf{r}, t)$ in eigenfunctions of $H_{\rm R}$ with time-dependent coefficients $a_\nu(t)$:

$$\psi(\mathbf{r}, t) = \sum_\nu a_\nu(t)\varphi_\nu(\mathbf{r})e^{-i\epsilon_\nu t/\hbar}.$$

For the expansion coefficients we have the initial condition $a_\nu(t = 0) = \langle\varphi_\nu|\phi_0\rangle$, because $\psi(\mathbf{r}, t = 0) = \phi_0(\mathbf{r})$. It is therefore useful to introduce new coefficients $c_\nu(t)$ defined by

$$a_\nu(t) = \langle\varphi_\nu|\phi_0\rangle e^{-i(E_0-\epsilon_\nu)t/\hbar} + c_\nu(t)$$

and the initial condition $c_\nu(t) = 0$. We can now express the wave function $\psi(\mathbf{r}, t)$ with these coefficients as

$$\psi(\mathbf{r}, t) = |\phi_0\rangle\, e^{-iE_0t/\hbar} + \sum_\nu |\varphi_\nu\rangle\, e^{-i\epsilon_\nu t/\hbar}c_\nu(t).$$

The equations for determining the coefficients $c_\nu(t)$ follow from the time-dependent Schrödinger equation with the hamiltonian H, i.e., from

$$\left(i\hbar\frac{\partial}{\partial t} - H\right)\psi(\mathbf{r}, t) = 0,$$

if the above expansion of the wave function is inserted. The result is the coupled inhomogeneous system of differential equations

$$i\hbar\frac{\partial}{\partial t}c_\mu(t) = t_{\mu,0}e^{-i(E_0-\epsilon_\mu)t/\hbar} + \Delta_\mu c_\mu(t) + \sum_{\nu(\neq\mu)} v_{\mu\nu}c_\nu(t)e^{-i(\epsilon_\nu-\epsilon_\mu)t/\hbar}$$

with the matrix elements

$$\begin{aligned}\Delta_\mu &= \langle\varphi_\mu|\, H-\epsilon_\mu\,|\varphi_\mu\rangle\\ t_{\mu\lambda} &= \langle\varphi_\mu|\, H-E_0\,|\phi_\lambda\rangle\\ v_{\mu\nu} &= \langle\varphi_\mu|\, H-\epsilon_\nu\,|\varphi_\nu\rangle\,.\end{aligned}$$

With the transformation

$$c_\mu(t)=\tilde{c}_\mu(t)e^{-i\Delta_\mu t/\hbar}$$

this system of equations simplifies to

$$i\hbar\frac{\partial}{\partial t}\tilde{c}_\mu(t)=t_{\mu,0}e^{-i(E_0-\epsilon_\mu-\Delta_\mu)t/\hbar}+\sum_{\nu(\neq\mu)}v_{\mu\nu}\tilde{c}_\nu(t)e^{-i(\epsilon_\nu-\epsilon_\mu)t/\hbar}.$$

By integration we obtain the system of integral equations

$$\begin{aligned}\tilde{c}_\mu(t)=\frac{i}{\hbar}t_{\mu,0}e^{-i(E_0-\epsilon_\mu-\Delta_\mu)t/2\hbar}\frac{\sin\left[(\epsilon_\mu+\Delta_\mu-E_0)t/2\hbar\right]}{-(\epsilon_\mu+\Delta_\mu-E_0)/2\hbar}\\ -\frac{i}{\hbar}\sum_\nu v_{\mu\nu}\int_0^t dt'\tilde{c}_\nu(t')e^{-i(\epsilon_\nu-\epsilon_\mu)t'/\hbar}.\end{aligned}$$

So far we have proceeded without making any approximations.

Now we treat the time evolution of the coefficients $\tilde{c}(t)$ using time-dependent perturbation theory in lowest order, taking the tunneling coupling matrix elements $t_{\mu,0}$ and $v_{\mu\nu}$ as small parameters. In lowest order we can neglect the second term on the right-hand side which contains the time integral, because it produces terms of higher order in the matrix elements. In this approximation we have

$$\tilde{c}_\mu(t)=\frac{i}{\hbar}t_{\mu,0}e^{-i(E_0-\epsilon_\mu-\Delta_\mu)t/2\hbar}\frac{\sin\left[(\epsilon_\mu+\Delta_\mu-E_0)t/2\hbar\right]}{-(\epsilon_\mu+\Delta_\mu-E_0)/2\hbar}.$$

The probability that the electron is found in state μ on the right side at time t is then given by

$$|\tilde{c}_\mu(t)|^2=\frac{1}{\hbar^2}\,|t_{\mu,0}|^2\,\frac{\sin^2\left[(\epsilon_\mu+\Delta_\mu-E_0)t/2\hbar\right]}{\left[(\epsilon_\mu+\Delta_\mu-E_0)/2\hbar\right]^2}.$$

We are now interested in the transition rate from the state ϕ_0 on the left side into the state φ_μ on the right side, i.e., we are interested in the time derivative

$$\frac{d\,|\tilde{c}_\mu(t)|^2}{dt}=\frac{2}{\hbar^2}\,|t_{\mu,0}|^2\,\frac{\sin\left[(\epsilon_\mu+\Delta_\mu-E_0)t/\hbar\right]}{(\epsilon_\mu+\Delta_\mu-E_0)/\hbar}.$$

For large times t, the fraction forming the last factor of this expression is significantly different from zero only if $\epsilon_\mu+\Delta_\mu-E_0$ is very close to zero. Using the expressions for the Dirac delta function in Appendix C we find

$$W_{\mu,0}=\lim_{t\to\infty}\frac{d\,|\tilde{c}_\mu(t)|^2}{dt}=\frac{2\pi}{\hbar}\,|t_{\mu,0}|^2\,\delta\left(\epsilon_\mu+\Delta_\mu-E_0\right). \qquad (12.2)$$

This result for the tunneling rate corresponds to Fermi's golden rule which is well known from time-dependent perturbation theory. Here the tunneling matrix element $t_{\mu,0}$ appears in the place of the matrix element of the potential perturbation. The delta function makes sure that only elastic tunneling processes take place in this order of perturbation theory. The matrix element Δ_μ describes the energy shift of level μ in the right contact which is caused by the tunneling coupling to the left contact (the so-called self-energy shift). The finite tunneling rate through the barrier leads to a finite lifetime of any state on each side of the barrier which is characterized by the square of the tunneling matrix element.

12.3 Tunneling current in a noninteracting system

The tunneling current that arises if a voltage is applied between the two electronic systems on the left and right sides of the tunneling barrier can be described using the same ideas that we used within the Landauer–Büttiker description of transport. The occupation of states on the left is given by the Fermi–Dirac distribution function $f_{\mathrm{L}}(E)$, the occupation of states on the right by $f_{\mathrm{R}}(E)$. The current can then be written as

$$I = -|e| \sum_{\mu,\nu} \left\{ f_{\mathrm{L}}(E_\mu) W_{\nu,\mu} [1 - f_{\mathrm{R}}(\epsilon_\nu)] - f_{\mathrm{R}}(\epsilon_\nu) W_{\mu,\nu} [1 - f_{\mathrm{L}}(E_\mu)] \right\},$$

which describes the difference between the current caused by electrons moving from left to right and those moving from right to left. This expression simplifies to

$$I = -|e| \sum_{\mu,\nu} W_{\nu,\mu} \left\{ f_{\mathrm{L}}(E_\mu) - f_{\mathrm{R}}(\epsilon_\nu) \right\},$$

if we use the property that $W_{\mu,\nu} = W_{\nu,\mu}$ as a result of time reversal symmetry. Inserting the expression for the transition matrix elements, this leads to

$$I = -|e| \frac{2\pi}{\hbar} \sum_{\mu,\nu} |t_{\nu,\mu}|^2 \, \delta(E_\mu - \epsilon_\nu) \left\{ f_{\mathrm{L}}(E_\mu) - f_{\mathrm{R}}(\epsilon_\nu) \right\}.$$

It can be rewritten as the integral over contributions of different energies via

$$I = -\frac{|e|}{h} \int dE \underbrace{(2\pi)^2 \sum_{\mu,\nu} |t_{\nu,\mu}|^2 \, \delta(E - \epsilon_\nu) \delta(E - E_\mu)}_{\mathcal{D}_{\mathrm{T}}(E)} \left\{ f_{\mathrm{L}}(E) - f_{\mathrm{R}}(E) \right\},$$

where $\mathcal{D}_{\mathrm{T}}(E)$ is called the *tunneling density of states*. If in a particular system all pairs of states (μ, ν) at a given energy E have about the same transition probability $\mathcal{T}(E) = |t_{\mu,\nu}|^2$, then the tunneling density of states can be simplified to

$$\mathcal{D}_{\mathrm{T}}(E) = (2\pi)^2 \mathcal{T}(E) \mathcal{D}_{\mathrm{L}}(E) \mathcal{D}_{\mathrm{R}}(E),$$

where the $\mathcal{D}_{\mathrm{L/R}}(E)$ are the densities of states in the left/right electron reservoir.

Tunneling spectroscopy. With the above simplified expression for the tunneling density of states we obtain the expression for the tunneling current

$$I = -\frac{(2\pi)^2|e|}{h}\int dE\mathcal{T}(E)\mathcal{D}_{\mathrm{L}}(E)\mathcal{D}_{\mathrm{R}}(E)\left\{f_{\mathrm{L}}(E) - f_{\mathrm{R}}(E)\right\}.$$

If the applied voltage $V_{\mathrm{SD}} = -(\mu_{\mathrm{R}} - \mu_{\mathrm{L}})/|e|$ between the left and the right reservoir is large compared to temperature, we may approximate the integral by its zero temperature expression

$$I = -\frac{(2\pi)^2|e|}{h}\int_{\mu_{\mathrm{R}}}^{\mu_{\mathrm{L}}} dE\mathcal{T}(E)\mathcal{D}_{\mathrm{L}}(E)\mathcal{D}_{\mathrm{R}}(E).$$

The derivative of the current with respect to V_{SD} is then given by

$$\frac{dI}{dV_{\mathrm{SD}}} = -\frac{(2\pi)^2|e|}{h}\mathcal{T}(\mu_{\mathrm{R}} + |e|V_{\mathrm{SD}})\mathcal{D}_{\mathrm{L}}(\mu_{\mathrm{R}} + |e|V_{\mathrm{SD}})\mathcal{D}_{\mathrm{R}}(\mu_{\mathrm{R}} + |e|V_{\mathrm{SD}}).$$

The derivative of the tunneling current reflects not only the energy dependence of the transmission probability $\mathcal{T}(E)$, but also that of the densities of states in the two reservoirs L and R. If the transmission is (over a certain energy range) independent of energy, then finite source–drain bias tunneling spectroscopy can be employed for a measurement of the joint density of states of the two reservoirs. If, in addition, the density of states of, say, the left reservoir is energy independent, then the density of states of the right reservoir can be directly measured. The latter situation is frequently given in the case of a scanning tunneling microscope, where a tunneling current flows from the apex of a metallic tip through a vacuum barrier into the (conducting) surface of interest. The density of states of the metallic tip may be regarded as constant over the energy range of interest, and the same can be assumed for the transmission, which has an exponential dependence on the tip–surface separation d, i.e., $\mathcal{T} \propto \exp(-2\kappa d)$. In this situation, the (local) surface density of states of the sample under investigation can be measured with tunneling spectroscopy.

The derivative of $I(V_{\mathrm{SD}})$ can be conveniently measured in practice by adding a low-frequency alternating voltage $V_{\mathrm{SD}}^{(\mathrm{AC})}$ of small amplitude to the source–drain voltage $V_{\mathrm{SD}}^{(\mathrm{DC})}$. The resulting alternating current $I^{(\mathrm{AC})}$ measured with lock-in techniques and normalized to $V_{\mathrm{SD}}^{(\mathrm{AC})}$ is then a good approximation to dI/dV_{SD}.

Small bias measurements. At small bias voltage V_{SD}, we can expand the difference of the Fermi–Dirac distribution functions according to eq. (11.7) and obtain the linear response expression for the conductance

$$G = \frac{|e|^2}{h}\int dE\mathcal{D}_{\mathrm{T}}(E)\left(\frac{\partial f}{\partial E}\right). \tag{12.3}$$

As a result we can say that the low temperature conductance probes the tunneling density of states at the Fermi energy.

12.4 Transfer hamiltonian

Often, the perturbative theoretical description of quantum tunneling is described on the basis of the so-called *transfer hamiltonian* which couples the states in the two leads. The hamiltonian for the whole system is then written in second quantization as

$$H = H_{\mathrm{L}} + H_{\mathrm{t}} + H_{\mathrm{R}},$$

where $H_{\mathrm{L/R}}$ is the hamiltonian of the left/right lead, and the transfer hamiltonian H_{t} is in second quantization given by

$$H_{\mathrm{t}} = \sum_{k,\kappa} \left(t_{\kappa k} b_{\kappa}^{\dagger} a_k + t_{\kappa k}^{\star} a_k^{\dagger} b_{\kappa} \right).$$

Here $b_{\kappa}^{\dagger}$ creates an electron in the right lead in state κ, whereas a_k deletes an electron in state k in the left lead. The $t_{\kappa k}$ are the tunneling matrix elements between the two contacts that we have introduced above. If H_{transfer} is treated in lowest order perturbation theory, Fermi's golden rule result is recovered.

Further reading

- Landauer–Büttiker formalism: Datta 1997; Beenakker and van Houten 1991.
- Paper: Bardeen 1961.

Exercises

(12.1) Discuss under which assumptions the Landauer–Büttiker result for the conductance in eq. (11.13) can be recovered from eq. (12.3).

Multiterminal systems

13

13.1 Generalization of conductance: conductance matrix

Now we consider the more general case of nanostructures with not just two, but n contacts. This situation is schematically shown in Fig. 13.1. In this generalized situation, the linear expansion of the current in terms of the voltages applied to the reservoirs [i.e., the generalization of Ohm's law in eq. (10.2)] leads to the matrix equation

$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \\ \vdots \\ I_n \end{pmatrix} = \begin{pmatrix} G_{11} & G_{12} & G_{13} & \dots & G_{1n} \\ G_{21} & G_{22} & G_{23} & \dots & G_{2n} \\ G_{31} & G_{32} & G_{33} & \dots & G_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ G_{n1} & G_{n2} & G_{n3} & \dots & G_{nn} \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ \vdots \\ V_n \end{pmatrix}. \qquad (13.1)$$

Here, the matrix of the conductance coefficients G_{ij} is the generalization of the conductance G. Two very fundamental considerations lead to relations between the conductance coefficients which have to be obeyed by all physically acceptable conductance matrices.

Consequence of the conservation of charge. In a transport experiment, electric charge is neither created nor destroyed: the charge is a conserved quantity. In electrodynamics this fact is often expressed by writing down the continuity equation $\frac{\partial \rho}{\partial t} + \nabla \mathbf{j} = 0$ where ρ is the charge density, and $\mathbf{j}$ is the electrical current density. In the case of stationary, time-independent problems, such as conductance measurements at zero frequency, the time derivative of the charge density is zero. As a consequence, the divergence of the current density must be zero. This implies that all currents entering and leaving the structure have to sum up to zero. This is Kirchhoff's current law that has to be obeyed for arbitrary applied voltages. Therefore, the conductance coefficients in each column of the conductance matrix fulfill the sum rule

$$\sum_{i=1}^{n} G_{ij} = 0. \qquad (13.2)$$

No transport currents without voltage differences between contacts. The second relation between conductance coefficients is obtained

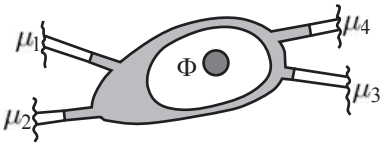


Fig. 13.1 Semiconductor nanostructure with four contacts consisting of perfect wires each connected to a reservoir with the electrochemical potential μ_i ($i = 1, 2, 3, 4$). (Reprinted with permission from Buttiker, 1986. Copyright 1986 by the American Physical Society.)

from the requirement that no currents will flow into or out of the structure, if all voltages applied to the contacts are the same. This leads to the sum rule

$$\sum_{j=1}^{n} G_{ij} = 0 \tag{13.3}$$

for the rows of the conductance matrix.

Using the two above sum rules we can show that the stationary currents in any nanostructure depend only on voltage differences between contacts:

$$I_i = -\sum_{\substack{j \\ j\neq i}} G_{ij}(V_i - V_j) = G_{ii}V_i + \sum_{\substack{j \\ j\neq i}} G_{ij}V_j = \sum_j G_{ij}V_j. \tag{13.4}$$

13.2 Conductance and transmission: Landauer–Büttiker approach

In the same way as for structures with two contacts, also in the general case of many terminals the conductance matrix can be related to the transmission probabilities of quantum states from one contact to the other. In order to show this, we again assume that the quantum mechanical many-particle problem has been solved for the thermodynamic equilibrium situation (i.e., $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots$), for example, using the Hartree approximation. We allow a homogeneous external magnetic field $\mathbf{B}$ to penetrate the structure.

Corresponding to our previous considerations for the quantum point contact, we write the asymptotic form of the scattering states in the leads which are assumed to be perfect wires as

$$\psi_m^\alpha(\mathbf{r}) = \frac{1}{\sqrt{L}} \begin{cases} \chi_m^\alpha(y,z)e^{ik_m^\alpha x} + \sum_n r_{nm}^\alpha \chi_n^\alpha(y,z)e^{-ik_n^\alpha x} & \text{for } x \to -\infty \\ \sum_{\beta,n} t_{nm}^{\beta\alpha}\chi_n^\beta(y,z)e^{ik_n^\beta x} & \text{for } x \to +\infty \end{cases}$$

Here, $t_{nm}^{\beta\alpha}$ describes the transmission amplitude from mode m in lead α into mode n in lead β. Correspondingly, r_{nm}^α is the reflection amplitude from mode m into mode n in the same lead α. The wave functions $\chi_m^\alpha(y,z)$ are the lateral modes in lead α. These scattering states describe a scattering experiment in which a particle is incident from mode m of lead α and scattered into any other mode in any other lead.

The above scattering state represents the current contribution

$$I_{\alpha\alpha} = -g_s \cdot \frac{|e|}{h} \int dE\,[N_\alpha(E) - \mathcal{R}_\alpha(E)] \cdot f_\alpha(E)$$

in lead α, where the number of modes at energy E in lead α is

$$N_\alpha(E) = \sum_{\substack{m \\ m\in\alpha}} 1,$$

and the reflection back into lead α is

$$\mathcal{R}_\alpha(E) = \sum_{\substack{n,m \\ n,m\in\alpha}} \frac{k_n^\alpha(E)}{k_m^\alpha(E)} \left|r_{nm}^\alpha(E)\right|^2 .$$

In addition, there are currents in lead α that arise from transmission from other contacts. An arbitrary contact $\beta \neq \alpha$ contributes the current

$$I_{\alpha\beta} = g_s \frac{|e|}{h} \int dE \mathcal{T}_{\alpha\beta}(E) \cdot f_\beta(E),$$

where we have defined the transmission $\mathcal{T}_{\alpha\beta}$ from lead β into lead α as

$$\mathcal{T}_{\alpha\beta}(E) = \sum_{\substack{n \\ n\in\alpha}} \sum_{\substack{m \\ m\in\beta}} \frac{k_n^\alpha(E)}{k_m^\beta(E)} \left|t_{nm}^{\alpha\beta}(E)\right|^2 . \tag{13.5}$$

The total current flowing in lead α is the incoming current minus the reflected current minus all the currents transmitted into lead α from other contacts. This results in

$$I_\alpha = -g_s \frac{|e|}{h} \int dE \Bigg\{ \left[N_\alpha(E) - \mathcal{R}_\alpha(E)\right] f_\alpha(E) - \sum_{\substack{\beta \\ \beta\neq\alpha}} \mathcal{T}_{\alpha\beta}(E) f_\beta(E) \Bigg\}. \tag{13.6}$$

Thermodynamic equilibrium. In the case $f_\alpha(E) = f_\beta(E)$, i.e., if all contacts are in thermodynamic equilibrium, no current will flow in the system and all $I_\alpha = 0$. It follows that the transmission and reflection probabilities fulfill the condition

$$N_\alpha(E) = \mathcal{R}_\alpha(E) + \sum_\beta \mathcal{T}_{\alpha\beta}(E). \tag{13.7}$$

It is the generalization of the condition that in a system with two single-moded leads reflection and transmission probabilities add to one ($\mathcal{R} + \mathcal{T} = 1$).

13.3 Linear response: conductance and transmission

For obtaining the linear response we expand the Fermi distribution functions $f_\beta(E)$ in eq. (13.6) for small differences $\mu_\beta - \mu_\alpha$. The result is

$$\begin{aligned} I_\alpha &= g_s \cdot \frac{|e|}{h} \sum_\beta \int dE \mathcal{T}_{\alpha\beta}(E) \left(\frac{\partial f_\alpha(E)}{\partial \mu}\right) (\mu_\beta - \mu_\alpha) \\ &= \sum_\beta G_{\alpha\beta} \cdot \left(\frac{\mu_\alpha}{-|e|} - \frac{\mu_\beta}{-|e|}\right). \end{aligned}$$

Denoting the voltage differences between contacts α and β as $V_\alpha - V_\beta$, we obtain in agreement with eq. (13.4)

$$I_\alpha = -\sum_{\substack{\beta \\ \beta\neq\alpha}} G_{\alpha\beta}\cdot(V_\alpha - V_\beta)\,. \tag{13.8}$$

Here we have introduced the off-diagonal elements of the conductance matrix in eq. (13.1)

$$G_{\alpha\beta} = -g_s\frac{e^2}{h}\int dE\mathcal{T}_{\alpha\beta}(E)\left(\frac{\partial f_\alpha(E)}{\partial\mu}\right). \tag{13.9}$$

Equations (13.5), (13.8) and (13.9) are the basis of the *Landauer–Büttiker formalism* for the calculation of the electrical conductance from the asymptotic form of the scattering states. These equations contain the special case of two-terminal structures such as the quantum point contact.

13.4 The transmission matrix

For a sample with n contacts, equation (13.8) can also be written in the matrix form

$$\begin{pmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \end{pmatrix} = \frac{e^2}{h}\begin{pmatrix} N_1-\mathcal{R}_1 & -\mathcal{T}_{12} & \dots & -\mathcal{T}_{1n} \\ -\mathcal{T}_{21} & N_2-\mathcal{R}_2 & \dots & -\mathcal{T}_{2n} \\ \vdots & \vdots & & \vdots \\ -\mathcal{T}_{n1} & -\mathcal{T}_{n2} & \dots & N_n-\mathcal{R}_n \end{pmatrix}\begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{pmatrix}. \tag{13.10}$$

Here, N_α denotes the number of modes in lead α, and the $\mathcal{R}_\alpha$ are the reflection probabilities. The transmission probabilities are here defined as

$$\mathcal{T}_{\alpha\beta} = g_s\int dE\;\mathcal{T}_{\alpha\beta}(E)\left(\frac{\partial f_\alpha(E)}{\partial\mu}\right)$$

and according to eq. (13.7)

$$N_\alpha - \mathcal{R}_\alpha = \sum_{\beta,\beta\neq\alpha}\mathcal{T}_{\alpha\beta}.$$

This equation ensures that no current flows when all contacts are at the same voltage, in complete agreement with eq. (13.3).

If we set the voltage V_α to 1 V and all other voltages to zero, charge conservation [eq. (13.2)] requires

$$N_\alpha - \mathcal{R}_\alpha = \sum_{\alpha,\alpha\neq\beta}\mathcal{T}_{\alpha\beta}.$$

When the Landauer–Büttiker formalism is used for calculating currents and voltages in an experiment, current and voltage contacts have to be distinguished. For the latter, the net current is zero, because

the connected voltmeter will have a very high internal resistance. The measured voltage on such a voltage terminal is then

$$V_\alpha = \sum_{\beta(\neq\alpha)} \frac{\mathcal{T}_{\alpha\beta}}{\sum_{\gamma(\neq\alpha)} \mathcal{T}_{\alpha\gamma}} V_\beta, \tag{13.11}$$

the average of the voltages on all other contacts transmitting into it weighted by the relative transmission strength.

Also relevant for the application of eq. (13.10) to experimental problems is the fact that not all rows of this matrix equation are linearly independent. This is seen, for example, by adding the first $n-1$ equations and using the column sum rules. The resulting equation is equal to the nth equation multiplied by -1. As a consequence, one equation of the system can be omitted. This is usually combined with the freedom of the choice of the voltage zero which allows us to regard one of the n voltages to be the zero voltage reference. If we chose $V_\alpha = 0$ and omit equation α, the problem reduces to a system of $n-1$ linear equations with $n-1$ unknown quantities.

13.5 S-matrix and T-matrix

Using the concept of S- and T-matrices, phase-coherent mesoscopic model systems can be described. Below we will introduce these matrices and show their most important properties.

S-matrix. The S-matrix, or scattering matrix, describes the relation between the amplitudes of states impinging onto a mesoscopic structure from contact leads, and the amplitudes of states reflected from, or transmitted through the structure into contact leads. This is schematically depicted in Fig. 13.2. The size of the matrix is given by the total number of modes in all leads connecting to the device. For example, if there are three leads, two of which support only one mode, but one carries two modes, then the S-matrix will be 4×4. The elements of the S matrix are the quantum mechanical probability amplitudes for transmission between any pair of modes.

The amplitudes $\tilde{b}_m^\alpha$ of outgoing and $\tilde{a}_n^\alpha$ of incoming waves are linearly related, i.e.,

$$\tilde{b}_m^\alpha = \sum_{\beta n, \beta\neq\alpha} t_{mn}^{\alpha\beta} \tilde{a}_n^\beta + \sum_n r_{mn}^\alpha \tilde{a}_n^\alpha.$$

We can write this relation in matrix form by taking the $\tilde{a}_n^\alpha$ as the components of a vector $\mathcal{A}'$ and the corresponding $\tilde{b}_m^\alpha$ as the components of a vector $\mathcal{B}'$, giving

$$\mathcal{B}' = \tilde{S}\mathcal{A}',$$

where the matrix $\tilde{S}$ contains the transmission and reflection coefficients. With this definition, the particle current represented by the amplitude $\tilde{a}_n^\alpha$ is given by

$$\frac{\hbar k_n^\alpha}{m^\star} \cdot \frac{1}{L} \cdot |\tilde{a}_n^\alpha|^2 .$$

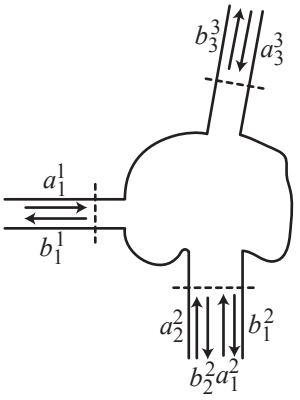


Fig. 13.2 Ingoing probability amplitudes a_n^α impinging onto a mesoscopic system and outgoing amplitudes b_n^α. The index α labels a particular lead, n numbers a particular mode in lead α.

A corresponding expression is valid for $\tilde{b}_n^\alpha$. It is convenient and common practice to introduce current amplitudes a_n^α and b_n^α which, when squared, directly give the current. We therefore define

$$a_n^\alpha := \sqrt{\frac{\hbar k_n^\alpha}{m^\star}} \cdot \tilde{a}_n^\alpha.$$

Again we take the a_n^α to be components of a vector $\mathcal{A}$ and the b_n^α as components of $\mathcal{B}$. Then the above matrix equation reads

$$\mathcal{B} = S \cdot \mathcal{A}$$

with S being the so-called scattering matrix (S-matrix) having elements

$$s_{nm}^{\alpha\beta} = \sqrt{\frac{k_n^\alpha}{k_m^\beta}}\tilde{S}_{nm}^{\alpha\beta} = \sqrt{\frac{k_n^\alpha}{k_m^\beta}} \begin{cases} r_{nm}^\alpha & \text{for } \alpha = \beta \\ t_{nm}^{\alpha\beta} & \text{for } \alpha \neq \beta \end{cases} .$$

Using these elements of the S-matrix we can express the transmission probability $\mathcal{T}_{\alpha\beta}(E)$ in eq. (13.5) in the more elegant form

$$\mathcal{T}_{\alpha\beta}(E) = \sum_{\substack{n \\ n\in\alpha}} \sum_{\substack{m \\ m\in\beta}} \left| s_{nm}^{\alpha\beta}(E) \right|^2 . \tag{13.12}$$

and the reflection $\mathcal{R}_\alpha(E)$ as

$$\mathcal{R}_\alpha(E) = \sum_{\substack{n \\ n\in\alpha}} \sum_{\substack{m \\ m\in\alpha}} \left| s_{nm}^{\alpha\alpha}(E) \right|^2 . \tag{13.13}$$

Charge conservation and unitarity of the S-matrix. The S-matrix has a number of useful properties. From the conservation of charge (Kirchhoff's current law, continuity equation) which means

$$\sum_{n\alpha} |b_n^\alpha|^2 = \sum_{n\alpha} |a_n^\alpha|^2 ,$$

we find immediately that

$$S^\dagger S = 1,$$

where $(S^\dagger)_{nm}^{\alpha\beta} = S_{mn}^{\beta\alpha\,\star}$. The S-matrix is unitary. This means for $(\alpha n) = (\beta m)$:

$$\sum_{\gamma p} \left| S_{pn}^{\gamma\alpha} \right|^2 = \sum_{\gamma p, \gamma\neq\alpha} \frac{k_p^\gamma}{k_n^\alpha} \left| t_{pn}^{\gamma\alpha} \right|^2 + \sum_p \frac{k_p^\alpha}{k_n^\alpha} \left| r_{pn}^\alpha \right|^2 = 1,$$

and for $(\alpha n \neq \beta m)$

$$\sum_{\gamma p} k_p^\gamma (t_{pn}^{\gamma\alpha})^\star t_{pm}^{\gamma\beta} = 0,$$

where we have defined $t_{nm}^{\alpha\alpha} = r_{nm}^\alpha$.

Example: *S*-matrix for a wire with a single mode. As an example, we consider the simplest case of a system with only a single mode in two identical leads α and β with $k^\alpha = k^\beta$. The S-matrix is 2×2. The unitarity requirement for the S-matrix allows a matrix of the general form

$$S = \begin{pmatrix} \sqrt{\frac{1}{2} - \epsilon} \cdot e^{i\delta_a} & \sqrt{\frac{1}{2} + \epsilon} \cdot e^{i(\theta+\varphi)} \\ \sqrt{\frac{1}{2} + \epsilon} \cdot e^{i(\theta-\varphi)} & -\sqrt{\frac{1}{2} - \epsilon} \cdot e^{i(2\theta-\delta_a)} \end{pmatrix}, \tag{13.14}$$

where $-1/2 \le \epsilon \le 1/2$. The parameter ϵ determines the transmission and reflection probabilities, whereas δ_a, θ, and φ are scattering phases relevant for interference phenomena. The matrix elements of the S-matrix can be identified with the transmission and reflection coefficients:

$$\begin{aligned} r^\alpha &= \sqrt{\frac{1}{2} - \epsilon} \cdot e^{i\delta_a} \\ r^\beta &= -\sqrt{\frac{1}{2} - \epsilon} \cdot e^{i(2\theta-\delta_a)} \\ t^{\alpha\beta} &= \sqrt{\frac{1}{2} + \epsilon} \cdot e^{i(\theta+\varphi)} \\ t^{\beta\alpha} &= \sqrt{\frac{1}{2} + \epsilon} \cdot e^{i(\theta-\varphi)} \end{aligned}$$

Extreme cases are $\epsilon = 1/2$ (perfect transmission) and $\epsilon = -1/2$ (total reflection).

***T*-matrix.** While the concept of the S-matrix is very general, the T-matrix (transfer matrix) is convenient only for the description of a series connection of coherent conductors each having two leads. The T-matrix relates all in- and outgoing amplitudes on one side (L) of the structure with those on the other side (R). We take the $a_n^{\rm L}$ and $b_n^{\rm L}$ to be components of vector $\mathcal{L}$, and the $a_n^{\rm R}$ and $b_n^{\rm R}$ to be components of vector $\mathcal{R}$. Then the T-matrix of a structure is defined by

$$\mathcal{R} = T\mathcal{L}.$$

The elements of the T-matrix can be expressed in terms of those of the S-matrix. We show this for a structure in which only a single mode exists in the left lead (L) and the right lead (R). In this case, the S-matrix is

$$S = \begin{pmatrix} r^{\rm L} & t^{\rm LR} \\ t^{\rm RL} & r^{\rm R} \end{pmatrix},$$

such that

$$\begin{pmatrix} b^{\rm L} \\ b^{\rm R} \end{pmatrix} = S \begin{pmatrix} a^{\rm L} \\ a^{\rm R} \end{pmatrix}.$$

We solve this matrix equation for the coefficients in the right lead, i.e.,

$$\begin{pmatrix} b^{\rm R} \\ a^{\rm R} \end{pmatrix} = T \begin{pmatrix} a^{\rm L} \\ b^{\rm L} \end{pmatrix},$$

and find for the T-Matrix

$$T = \begin{pmatrix} (t^{\mathrm{LR}}t^{\mathrm{RL}} - r^{\mathrm{L}}r^{\mathrm{R}})/t^{\mathrm{LR}} & r^{\mathrm{R}}/t^{\mathrm{LR}} \\ -r^{\mathrm{L}}/t^{\mathrm{LR}} & 1/t^{\mathrm{LR}} \end{pmatrix}.$$

13.6 Time-reversal invariance and magnetic field

It can be shown that in the presence of a magnetic field $\mathbf{B}$ the S-matrix obeys

$$S(\mathbf{B}) = S^T(-\mathbf{B}).$$

In order to prove this relation, we consider Schrödinger's equation

$$\left\{ \frac{[-i\hbar\nabla + |e|\mathbf{A}(\mathbf{r})]^2}{2m^\star} + V(\mathbf{r}) \right\} \psi(\mathbf{r};\mathbf{B}) = E\psi(\mathbf{r};\mathbf{B}).$$

If we take the conjugate complex of this equation and reverse the direction of the magnetic field $\mathbf{B}$, we find

$$\left\{ \frac{[-i\hbar\nabla + |e|\mathbf{A}(\mathbf{r})]^2}{2m^\star} + V(\mathbf{r}) \right\} \psi^\star(\mathbf{r};-\mathbf{B}) = E\psi^\star(\mathbf{r};-\mathbf{B}).$$

It follows that

$$\psi^\star(\mathbf{r};-\mathbf{B}) = \psi(\mathbf{r};\mathbf{B}),$$

i.e., if $\psi(\mathbf{r};\mathbf{B})$ solves Schrödinger's equation for magnetic field $\mathbf{B}$, then a solution of the problem with magnetic field direction reversed can be obtained from this equation. We now apply this property to the asymptotic scattering states used in the Landauer–Büttiker description of electronic transport. Taking the conjugate complex of such a scattering state, the ingoing amplitudes $\mathcal{A}$ become the conjugate complex of the outgoing scattering states $\mathcal{B}^\star$. This corresponds to a reversal of the time direction. Therefore we have

$$\mathcal{B} = S(\mathbf{B})\mathcal{A} \Rightarrow \mathcal{B}^\star = S^\star(\mathbf{B})\mathcal{A}^\star$$

and

$$\mathcal{A}^\star = S(-\mathbf{B})\mathcal{B}^\star \Rightarrow \mathcal{B}^\star = S^{-1}(-\mathbf{B})\mathcal{A}^\star.$$

As a result,

$$S^{-1}(-\mathbf{B}) = S^\dagger(-\mathbf{B}) = S^\star(\mathbf{B}),$$

where we have used the unitarity of S. It follows that

$$S(\mathbf{B}) = S^T(-\mathbf{B}), \tag{13.15}$$

and the matrix elements of the S-matrix obey

$$s^{\alpha\beta}_{nm}(B) = s^{\beta\alpha}_{mn}(-B).$$

At zero magnetic field, the S-matrix is symmetric, as a result of time-reversal invariance of the problem, i.e.,

$$S^{-1}(\mathbf{B}=0) = S^T(\mathbf{B}=0).$$

From the above symmetry relations for the S-matrix we can deduce important symmetry relations for the conductance matrix and the transmission probabilities. For the latter we find from using eqs (13.12) and (13.15)

$$\mathcal{T}_{\alpha\beta}(E,\mathbf{B}) = \mathcal{T}_{\beta\alpha}(E,-\mathbf{B}) \quad (13.16)$$

$$N_\alpha(E,\mathbf{B}) - \mathcal{R}_\alpha(E,\mathbf{B}) = N_\alpha(E,-\mathbf{B}) - \mathcal{R}_\alpha(E,-\mathbf{B}). \quad (13.17)$$

As a result of eq. (13.9) we obtain the general symmetry relation

$$G_{\alpha\beta}(\mathbf{B}) = G_{\beta\alpha}(-\mathbf{B}). \quad (13.18)$$

This symmetry relation is known as the *generalized Onsager relation* for the conductance matrix.

13.7 Four-terminal resistance

As in the case of diffusive Drude transport, driving a current through the structure between two leads and measuring the voltage between two other leads avoids the contribution of the resistances of the ohmic contacts to the measurement results. The same is true for mesoscopic samples. We will therefore discuss the Landauer–Büttiker description of an arbitrary four-terminal measurement below.

The starting point is an arbitrary mesoscopic device with four contact leads as depicted in Fig. 13.3. The structure is exposed to a magnetic field $\mathbf{B}$. Following Buttiker 1986, we consider a current I_1 driven from terminal 1 to 3 and a current I_2 from 2 to 4. This is a slightly more general approach than needed, but it is more symmetric than the standard four-terminal situation in which, for example, $I_2 = 0$. The currents and voltages in the four terminals are related via the four equations

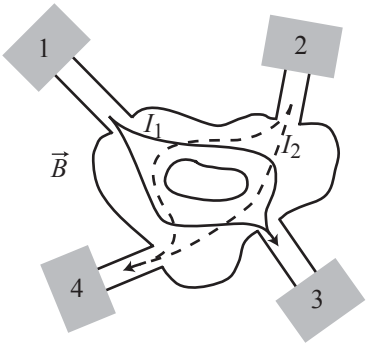


Fig. 13.3 Schematic representation of a four-terminal mesoscopic device in which a current I_1 flows from contact 1 to 3, and a current I_2 flows from contact 2 to 4.

$$I_1 = \frac{e^2}{h}[\mathcal{T}_{12}(V_1-V_2)+\mathcal{T}_{13}(V_1-V_3)+\mathcal{T}_{14}(V_1-V_4)] \quad (13.19)$$

$$I_2 = \frac{e^2}{h}[\mathcal{T}_{21}(V_2-V_1)+\mathcal{T}_{23}(V_2-V_3)+\mathcal{T}_{24}(V_2-V_4)] \quad (13.20)$$

$$-I_1 = \frac{e^2}{h}[\mathcal{T}_{31}(V_3-V_1)+\mathcal{T}_{32}(V_3-V_2)+\mathcal{T}_{34}(V_3-V_4)] \quad (13.21)$$

$$-I_2 = \frac{e^2}{h}[\mathcal{T}_{41}(V_4-V_1)+\mathcal{T}_{42}(V_4-V_2)+\mathcal{T}_{43}(V_4-V_3)]. \quad (13.22)$$

We now wish to express the currents in terms of the three voltage differences $V_1 - V_3$, $V_2 - V_4$, and $V_3 - V_4$. This can be accomplished by realizing that $V_1 - V_2 = (V_1 - V_3) - (V_2 - V_4) + (V_3 - V_4)$, $V_1 - V_4 = (V_1 - V_3) + (V_3 - V_4)$, and $V_2 - V_3 = (V_2 - V_4) - (V_3 - V_4)$. Using in

addition the sum rule (13.3) we obtain

$$\begin{aligned} I_1 &= \frac{e^2}{h}[(N_1-\mathcal{R}_1)(V_1-V_3)-\mathcal{T}_{12}(V_2-V_4) \\ &\quad +(\mathcal{T}_{12}+\mathcal{T}_{14})(V_3-V_4)] \qquad (13.23) \\ I_2 &= \frac{e^2}{h}[(N_2-\mathcal{R}_2)(V_2-V_4)-\mathcal{T}_{21}(V_1-V_3) \\ &\quad -(\mathcal{T}_{21}+\mathcal{T}_{23})(V_3-V_4)] \qquad (13.24) \\ -I_1 &= \frac{e^2}{h}[-\mathcal{T}_{31}(V_1-V_3)-\mathcal{T}_{32}(V_2-V_4)+(\mathcal{T}_{32}+\mathcal{T}_{34})(V_3-V_4)] \qquad (13.25) \\ -I_2 &= \frac{e^2}{h}[-\mathcal{T}_{42}(V_2-V_4)-\mathcal{T}_{41}(V_1-V_3)-(\mathcal{T}_{41}+\mathcal{T}_{43})(V_3-V_4)]. \qquad (13.26) \end{aligned}$$

Summing (13.23) and (13.25), or (13.24) and (13.26), we get, after using (13.2), the relation

$$\Sigma(V_3-V_4) = -(\mathcal{T}_{41}+\mathcal{T}_{21})(V_1-V_3)+(\mathcal{T}_{12}+\mathcal{T}_{32})(V_2-V_4) \qquad (13.27)$$

where $\Sigma := \mathcal{T}_{12}+\mathcal{T}_{14}+\mathcal{T}_{32}+\mathcal{T}_{34} = \mathcal{T}_{21}+\mathcal{T}_{41}+\mathcal{T}_{23}+\mathcal{T}_{43}$. If we now insert eq. (13.27) into eqs (13.23) and (13.24), we obtain a relation between the two currents I_1, I_2 and the two voltage differences V_1-V_3, V_2-V_4 which is of the form

$$\begin{pmatrix} I_1 \\ I_2 \end{pmatrix} = \begin{pmatrix} \alpha_{11} & -\alpha_{12} \\ -\alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} V_1-V_3 \\ V_2-V_4 \end{pmatrix}, \qquad (13.28)$$

where

$$\begin{aligned} \alpha_{11} &= \frac{e^2}{h}\frac{(N_1-\mathcal{R}_1)\Sigma-(\mathcal{T}_{41}+\mathcal{T}_{21})(\mathcal{T}_{12}+\mathcal{T}_{14})}{\Sigma} \qquad (13.29) \\ \alpha_{12} &= \frac{e^2}{h}\frac{\mathcal{T}_{12}\mathcal{T}_{34}-\mathcal{T}_{32}\mathcal{T}_{14}}{\Sigma} \qquad (13.30) \\ \alpha_{21} &= \frac{e^2}{h}\frac{\mathcal{T}_{43}\mathcal{T}_{21}-\mathcal{T}_{23}\mathcal{T}_{41}}{\Sigma} \qquad (13.31) \\ \alpha_{22} &= \frac{e^2}{h}\frac{(N_2-\mathcal{R}_2)\Sigma-(\mathcal{T}_{12}+\mathcal{T}_{32})(\mathcal{T}_{21}+\mathcal{T}_{23})}{\Sigma}. \qquad (13.32) \end{aligned}$$

As a consequence of the symmetry relations (13.16) and (13.17), we have the new symmetries $\Sigma(\mathbf{B}) = \Sigma(-\mathbf{B})$, and $\alpha_{nm}(\mathbf{B}) = \alpha_{mn}(-\mathbf{B})$ $(n,m \in \{0,1\})$.

It is now straightforward to derive two- and four-terminal resistances from eq. (13.28). For example, if we pass the current I_1 from contact 1 to 3 and measure the voltage between contacts 2 and 4, then $I_2 = 0$ and the four-terminal resistance is

$$R_{13,24} = \frac{V_2-V_4}{I_1} = \frac{\alpha_{21}}{\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21}}.$$

If we pass the current I_2 from contact 2 to 4 and measure the voltage between contacts 1 and 3, then $I_1 = 0$ and the four-terminal resistance is

$$R_{24,13} = \frac{V_1 - V_3}{I_2} = \frac{\alpha_{12}}{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}}.$$

The denominator of both resistance expressions is an even function of the magnetic field. However, the four-terminal resistance need not be an even function of the magnetic field, because the numerator does not have this symmetry. The two four-terminal resistances for which the roles of current and voltage leads are interchanged have the symmetry property

$$R_{13,24}(\mathbf{B}) = R_{24,13}(-\mathbf{B}).$$

This symmetry property is found in all four-terminal linear transport experiments. It implies that, at zero magnetic field, current and voltage terminals can be interchanged without changing the measured resistance.

Changing the direction of the current, or that of the voltage, changes the sign of the measured resistance. For example,

$$R_{13,24}(\mathbf{B}) = -R_{31,24}(\mathbf{B}) = -R_{13,42}(\mathbf{B}) = R_{31,42}(\mathbf{B}).$$

We now determine the two-terminal resistances from eq. (13.28) and find

$$\begin{aligned} R_{13,13} &= \frac{\alpha_{22}}{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}} \\ R_{24,24} &= \frac{\alpha_{11}}{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}}. \end{aligned}$$

The two-terminal resistances of a mesoscopic structure are therefore always even in magnetic field, i.e.,

$$\begin{aligned} R_{13,13}(\mathbf{B}) &= R_{13,13}(-\mathbf{B}) \\ R_{24,24}(\mathbf{B}) &= R_{24,24}(-\mathbf{B}). \end{aligned}$$

In general, there are six different ways in which a current can be passed between contacts in a four-terminal geometry. The above treatment covers two of these cases. The other four cases can be obtained by relabeling contacts. The general expression for the four-terminal resistance is

$$R_{mn,kl} = \frac{V_k - V_l}{I_{m\to n}} = \frac{h}{e^2}\frac{\mathcal{T}_{ln}\mathcal{T}_{km} - \mathcal{T}_{kn}\mathcal{T}_{lm}}{D/\Sigma},$$

where the quantity in the denominator,

$$\begin{aligned} D = {} & [(N_1 - \mathcal{R}_1)\Sigma - (\mathcal{T}_{41} + \mathcal{T}_{21})(\mathcal{T}_{12} + \mathcal{T}_{14})] \\ & \times [(N_2 - \mathcal{R}_2)\Sigma - (\mathcal{T}_{12} + \mathcal{T}_{32})(\mathcal{T}_{21} + \mathcal{T}_{23})] \\ & - (\mathcal{T}_{12}\mathcal{T}_{34} - \mathcal{T}_{32}\mathcal{T}_{14})(\mathcal{T}_{43}\mathcal{T}_{21} - \mathcal{T}_{23}\mathcal{T}_{41}), \end{aligned}$$

is the same for all configurations and even in magnetic field.

13.8 Ballistic transport experiments in open systems

The Landauer–Büttiker formalism is, for example, very successful in describing ballistic transport experiments. These are experiments on nanostructures with characteristic dimensions much smaller than the elastic mean free path of the charge carriers. The quantization of the conductance in a quantum point contact was our first introductory example for a ballistic nanostructure having only two terminals. Below we will describe ballistic experiments on structures with more than two contacts. Four-terminal structures are of particular relevance for experiments. The systems we will consider here are known as open or strongly coupled systems. Although small on the scale of the elastic mean free path, they are strongly coupled to voltage and current contacts through resistances that are small or of the order of the conductance quantum. This is in contrast to structures that are only very weakly coupled, for example, to a source and a drain contact, through tunneling contacts. Examples of the latter type would be the tunneling contact between the conductive tip of a scanning tunneling microscope and a conducting surface, a quantum point contact which is strongly pinched off, or a quantum dot structure.

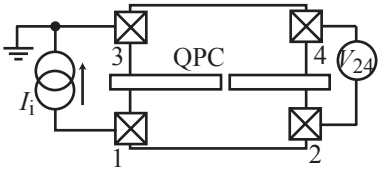


Fig. 13.4 Schematic picture of a four-terminal setup for measuring the conductance of a quantum point contact.

Four-terminal measurement of a quantum point contact. As a first example, we consider again a quantum point contact, but now measured in a four-terminal setup as shown schematically in Fig. 13.4. The current I_i is driven from terminal 1 to 3, the voltage V_c is measured between contacts 2 and 4. This example shows how we can simplify situations in which certain transmission functions are much bigger than others. In our example, the transmissions $\mathcal{T}_{12}$, $\mathcal{T}_{21}$, $\mathcal{T}_{34}$, and $\mathcal{T}_{43}$ can be expected to be much bigger than those between pairs of contacts on opposite sides of the QPC. In such cases we can introduce a small parameter, say ϵ, which quantifies the order of magnitude of a particular transmission. We would say, $\mathcal{T}_{12}$, $\mathcal{T}_{21}$, $\mathcal{T}_{34}$, and $\mathcal{T}_{43}$ are of order ϵ^0, whereas $\mathcal{T}_{13}$, $\mathcal{T}_{31}$, $\mathcal{T}_{14}$, $\mathcal{T}_{41}$, etc. are of order ϵ^1. The case $\epsilon = 0$ means that the QPC is fully pinched off.

In the experiment, the injected current I_i is given by the external current source. We now expand the voltage differences

$$V_i - V_j = V_{ij}^{(0)} + \epsilon V_{ij}^{(1)} + \epsilon^2 V_{ij}^{(2)} + \dots . \tag{13.33}$$

With this expansion in mind, we have a fresh look at eqs (13.19)–(13.22). Considering in eq. (13.20) with $I_2 = 0$ only terms of order ϵ^0, we find $V_{12}^{(0)} = 0$, i.e., $V_1 = V_2$ in lowest order. In the same way, eq. (13.22) leads in lowest order to $V_{34}^{(0)} = 0$, i.e., $V_3 = V_4$ in lowest order.

We extend our considerations to the order ϵ^1, because we see that eq. (13.19) is at least of this order. Considering the first order contribu-

tions in eqs (13.19) and (13.20) we find

$$\begin{aligned} I_\mathrm{i} &= \frac{e^2}{h}[\mathcal{T}_{12}^{(0)}V_{12}^{(1)} + (\mathcal{T}_{13}^{(1)} + \mathcal{T}_{14}^{(1)})V_{24}^{(0)}] \\ 0 &= -\mathcal{T}_{21}^{(0)}V_{12}^{(1)} + (\mathcal{T}_{23}^{(1)} + \mathcal{T}_{24}^{(1)})V_{24}^{(0)}. \end{aligned}$$

We have added the orders of the respective transmission probabilities as superscripts for clarity. It is now straightforward to determine $V_{12}^{(1)}$ from the second equation and insert it into the first. As a result we obtain in this first order approximation the inverse of the four-terminal resistance of the quantum point contact

$$\frac{1}{R_{13,24}} = \frac{I_i}{V_{24}^{(0)}} = \frac{e^2}{h}\left[\frac{\mathcal{T}_{12}}{\mathcal{T}_{21}}(\mathcal{T}_{23} + \mathcal{T}_{24}) + \mathcal{T}_{13} + \mathcal{T}_{14}\right].$$

At magnetic field $B = 0$ we have $\mathcal{T}_{12} = \mathcal{T}_{21}$ and the result simplifies to

$$\frac{1}{R_{13,24}} = \frac{e^2}{h}\left[\mathcal{T}_{23} + \mathcal{T}_{24} + \mathcal{T}_{13} + \mathcal{T}_{14}\right].$$

The sum in square brackets is the total transmission function through the QPC. It will be dominated by the quantized conductance and therefore increase in quantized steps of $2e^2/h$, given that the material has a single conduction band minimum (valence band maximum) with spin degeneracy.

Experimentally, the advantage of this four-terminal setup is that parasitic resistances arising at ohmic contacts do not play a role here. Although we have not included these in the above description, a two-terminal measurement $R_{13,13}$ would suffer from the addition of these contact resistances. In contrast, they do not appear in $R_{13,24}$, because there is no current flow through the voltage terminals 2 and 4 ($I_2 = 0$) as a result of the (infinite) internal impedance of the voltmeter.

Magnetic steering. Another simple structure consists of two quantum point contacts connected in series as schematically depicted in Fig. 13.5. Using the split-gate electrodes, regions of the two-dimensional electron gas below can be depleted and the number of modes in the quantum point contacts can be controlled. Four ohmic contacts connect the structure to the external world. Contact 1 is the source for quantum point contact 1 (QPC1), contact 3 is the drain contact of QPC1 and together with contact 2 at the same time the source of QPC2, contact 4 is the drain contact of QPC2. In the experiment contact 3 is grounded and a current I_i is injected from contact 1 via QPC 1 into contact 3. The measured quantity is the (collector) voltage V_c between contacts 4 and 2. There is a very low resistance connection between contacts 2 and 3.

The analysis of the general four-terminal resistance measurement can again be applied to this situation. As in the above example, simplifications arise because the transmission functions $\mathcal{T}_{\alpha\beta}$ differ strongly in magnitude. The quantities $\mathcal{T}_{23}$ and $\mathcal{T}_{32}$ will be of the order ϵ^0, followed

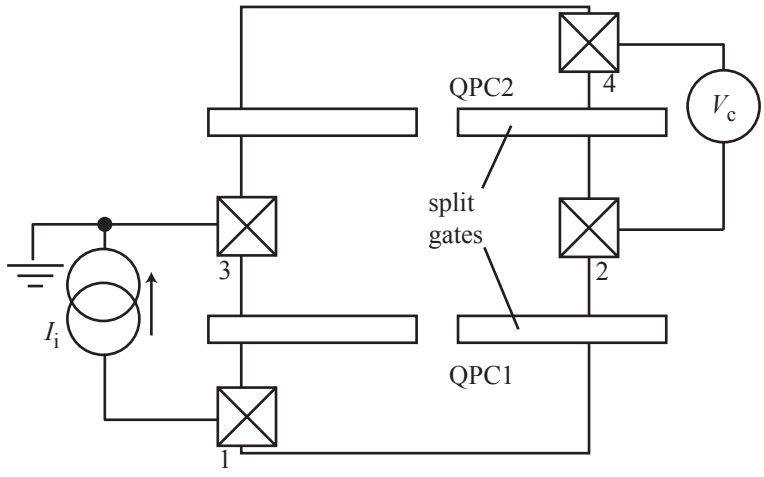


Fig. 13.5 Schematic representation of a setup for a magnetic steering experiment.

by the transmission functions through one of the QPCs ($\mathcal{T}_{12}$, $\mathcal{T}_{21}$, $\mathcal{T}_{13}$, $\mathcal{T}_{31}$, $\mathcal{T}_{24}$, $\mathcal{T}_{42}$, $\mathcal{T}_{34}$, $\mathcal{T}_{43}$) which are of the order ϵ^1. The transmissions $\mathcal{T}_{14}$ and $\mathcal{T}_{41}$ involving both QPCs will be of the order ϵ^2. In the experiment, the injected current I_i is given by the current source, and the voltages are again expanded in ϵ according to eq. (13.33).

Inserting this expansion in eq. (13.20) with $I_2 = 0$, we get in lowest order $V_{23}^{(0)} = 0$, i.e., $V_2 = V_3$, as expected from the highly conductive connection between the two contacts. Equation (13.19) is again at least of the order ϵ^1. We have to extend our considerations to this order, if we want to find the two- and four-terminal resistances. From eq. (13.22) we find in the order ϵ^1 the relation $V_{24}^{(0)} = 0$, i.e., $V_c = 0$. From eq. (13.19) we find in the same order the relation

$$I_i = \frac{e^2}{h}[\mathcal{T}_{12}^{(1)} + \mathcal{T}_{13}^{(1)}]V_{13}^{(0)}$$

between the current and the voltage between contacts 1 and 3, implying that the two-terminal resistance is given by

$$R_{13,13}^{-1} = \frac{I_i}{V_{13}^{(0)}} = \frac{e^2}{h}[N_1 - \mathcal{R}_1]$$

which is the resistance of QPC1. From eq. (13.20) we find the first order contribution $V_{23}^{(1)} = \mathcal{T}_{21}^{(1)}/\mathcal{T}_{23}^{(0)}V_{13}^{(0)}$. It expresses the fact that the finite transmission from contact 1 to 2 raises the voltage on terminal 2 compared to that on terminal 3 as soon as there is finite current flow [remember that the voltage on a voltage terminal will always be the average of the voltages on all other terminals weighted by the relative transmission from the respective terminal, eq. (13.11)].

Eventually we are interested in the lowest order contribution to $V_c = -V_{24}$. In order to obtain it, we have to look at eq. (13.22) in second order which reads

$$0 = -\mathcal{T}_{41}^{(2)}V_{14}^{(0)} - \mathcal{T}_{42}^{(1)}V_{24}^{(1)} - \mathcal{T}_{43}^{(1)}V_{34}^{(1)}.$$

For simplifying the third term we use $V_{34} = V_{24} - V_{23}$, and for the first term we employ $V_{14} = V_{24} + V_{13} - V_{23}$ which leads to $V_{14}^{(0)} = V_{13}^{(0)}$. As a result we get

$$V_{24}^{(1)} = \frac{-\mathcal{T}_{41}^{(2)} V_{13}^{(0)} + \mathcal{T}_{43}^{(1)} V_{23}^{(1)}}{\mathcal{T}_{42}^{(1)} + \mathcal{T}_{43}^{(1)}}.$$

The voltage V_{24} is a weighted average of $V_{13}^{(0)} = V_{12}^{(0)}$ and $V_{23}^{(1)}$.

Using the above results for $V_{13}^{(0)}$ and $V_{23}^{(1)}$ we obtain

$$-V_{\mathrm{c}} = V_{24}^{(1)} = -\frac{h}{e^2} \frac{\mathcal{T}_{41}^{(2)} \mathcal{T}_{23}^{(0)} - \mathcal{T}_{43}^{(1)} \mathcal{T}_{21}^{(1)}}{\mathcal{T}_{23}^{(0)} (\mathcal{T}_{42}^{(1)} + \mathcal{T}_{43}^{(1)})(\mathcal{T}_{12}^{(1)} + \mathcal{T}_{13}^{(1)})} I_{\mathrm{i}},$$

and we can write for the four-terminal resistance in this approximation

$$R_{13,42} = \frac{V_{\mathrm{c}}}{I_{\mathrm{i}}} = \frac{h}{e^2} \frac{\mathcal{T}_{41}\mathcal{T}_{23} - \mathcal{T}_{43}\mathcal{T}_{21}}{\mathcal{T}_{23}(N_4 - \mathcal{R}_4)(N_1 - \mathcal{R}_1)}. \tag{13.34}$$

The two bracketed expressions in the denominator are the conductances of the two individual QPCs. The two terms in the numerator split the resistance into two distinct parts: the first part is strongly sensitive to the ballistic transmission $\mathcal{T}_{41}$, whereas the second part is the product of two QPC transmission functions normalized to the large quantity $\mathcal{T}_{23}$. This second term can be made small compared to the first, if $\mathcal{T}_{23}$ is made very big (which can be achieved using an appropriate geometry). The idea of the magnetic steering experiment is now to tune $\mathcal{T}_{14}$ using a small magnetic field. If the two QPCs are properly aligned, $\mathcal{T}_{14}$ will be largest at zero magnetic field, whereas it will be strongly suppressed even at small magnetic fields, where the Lorentz force bends the beam of ballistic electrons ejected from QPC1 away from the QPC2 opening. The experimental result is the sharp maximum of $R_{13,42}$ at zero magnetic field shown in Fig. 13.6. At finite magnetic fields beyond the peak of the resistance, the second contribution to the resistance takes over and produces a small negative value of V_{c}.

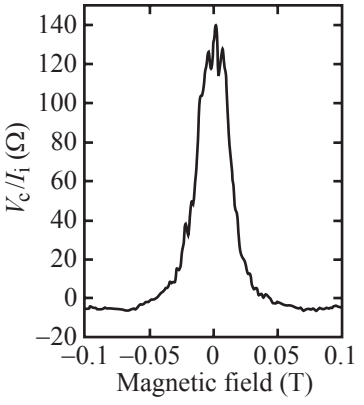


Fig. 13.6 Result of a magnetic steering experiment. Plotted is the nonlocal resistance $V_{\mathrm{c}}/I_{\mathrm{i}}$. (Reprinted with permission from Molenkamp *et al.*, 1990. Copyright 1990 by the American Physical Society.)

Magnetic focusing. A further example of a ballistic transport experiment with magnetic field is the so-called magnetic focusing experiment. It is similar to the magnetic steering experiment in that the structure has four contacts and two quantum point contact constrictions, but here the two QPCs are placed next to each other as depicted in Fig. 13.7. Certain values of the magnetic field lead to a cyclotron radius R_{c} that is an integer fraction of the separation d of the two slits, i.e.,

$$R_{\mathrm{c}} n = d \Rightarrow \frac{\hbar k_{\mathrm{F}}}{|e| B_n} n = d \Rightarrow B_n = \frac{\hbar k_{\mathrm{F}}}{|e| d} n \text{ with } n \text{ a positive integer.}$$

The description of the nonlocal resistance with the Landauer–Büttiker formalism is identical to that of the magnetic steering experiment. If the magnetic field fulfills the above condition, $\mathcal{T}_{41}$ is maximum. At the same time, $\mathcal{T}_{41} \approx 0$, if the focusing condition is not fulfilled. As in the magnetic steering geometry the four-terminal resistance is given by

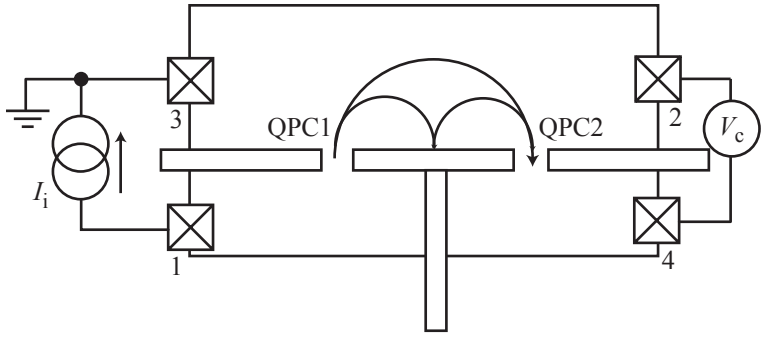


Fig. 13.7 Schematic setup of a magnetic focusing experiment.

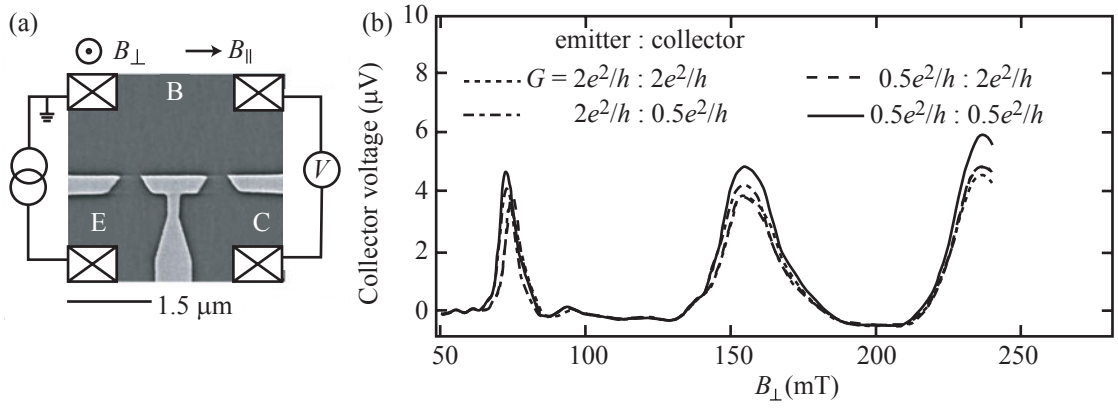


Fig. 13.8 (a) Image of the sample and schematic of the external circuit. (b) Collector voltage measured in a magnetic focusing geometry as a function of the magnetic field applied perpendicular to the plane of the two-dimensional electron gas. The heights of the resonances depend only weakly from the opening of the two quantum point contacts. (Reprinted with permission from Potok *et al.*, 2002. Copyright 2002 by the American Physical Society.)

eq. (13.34). Correspondingly, the voltage V_c will be maximum whenever $\mathcal{T}_{41}$ is maximum. We therefore expect equidistant maxima in the magnetic field dependence of $R_{13,42}$. Figure 13.8 shows the result of an experiment. It can be seen that the expected maxima exist. Between these maxima, the collector voltage V_c is slightly negative, but very small compared to the value at the maxima. Therefore, the expression of the resistance in this experiment can be simplified to

$$R_{13,42} = \frac{h}{e^2} \frac{\mathcal{T}_{41}}{(N_4 - \mathcal{R}_4)(N_1 - \mathcal{R}_1)}.$$

It is interesting to see in Fig. 13.8 that in this sample with very high mobility, the heights of the maxima do not depend very strongly on the opening of the quantum point contacts (Potok *et al.*, 2002). In this experiment, the injected current was $I = 1\,\mathrm{nA}$, and the height of $5\,\mu\mathrm{V}$ of

the measured resistance peaks corresponds therefore to a resistance of about $5\,\mathrm{k}\Omega$, or $0.2h/e^2$. This means that there is the empirical relation

$$\mathcal{T}_{41} = k(N_1 - \mathcal{R}_1)(N_4 - \mathcal{R}_4) \text{ on a focusing peak,} \tag{13.35}$$

with $k \approx 0.2$.

Quantum point contact spin filter. Using a quantum point contact, a spin filter can be realized (Potok *et al.*, 2003). For this purpose, the focusing geometry can be used as depicted in Fig. 13.8(a). The current is spin-polarized by the quantum point contact on the left by applying a strong magnetic field parallel to the electron gas. If we assume that the spin of electrons tunneling from the emitter E to the basis B is conserved, a fully spin-polarized current is expected if only one of the two spin states exists at the Fermi energy in the emitter. If both spin states exist at the Fermi energy they may have different transmission probabilities to the base, and the current is partially spin-polarized. On the right side of the structure there is a quantum point contact (analyzer) exposed to the same in-plane magnetic field as the emitter. The magnetic field lifts the degeneracy of the modes of both quantum point contacts, and conductance plateaus are observed at integer multiples of e^2/h (rather than those at $2e^2/h$ observed at zero in-plane field). This suggests that the quantum point contacts transmit only one spin direction at gate voltages where the conductance is below the first conductance plateau, and therefore act as spin filter devices.

For an analytic description of the setup we use the Landauer–Büttiker formalism. We define the spin selectivity of the collector quantum point contact to be

$$P_\mathrm{c} = \frac{G_{33}^{\uparrow} - G_{33}^{\downarrow}}{G_{33}^{\uparrow} + G_{33}^{\downarrow}},$$

and that of the emitter (here it is a quantum point contact, but it could also be a quantum billiard or a quantum dot)

$$P_\mathrm{e} = \frac{G_{11}^{\uparrow} - G_{11}^{\downarrow}}{G_{11}^{\uparrow} + G_{11}^{\downarrow}}.$$

The values of these polarizations range from -1 (only spin-down transmitted), via 0 (both spin orientations transmitted with the same probability) to $+1$ (only spin-up transmitted). In an ideal polarizer–analyzer geometry the output voltage V_c would be given by

$$V_\mathrm{c} \propto \frac{1}{2}(1 + P_\mathrm{e}P_\mathrm{c}) = \frac{G_{11}^{\uparrow}G_{33}^{\uparrow} + G_{11}^{\downarrow}G_{33}^{\downarrow}}{G_{11}G_{33}}.$$

Zero voltage is measured if P_e and P_c have the same magnitude but opposite signs. The maximum voltage is measured if both signs are the same and the magnitudes are one. If we compare this with the expression for the nonlocal resistance in a magnetic focusing experiment

$$\frac{V_c}{I} = \frac{G_{31}}{G_{11}G_{33}},$$

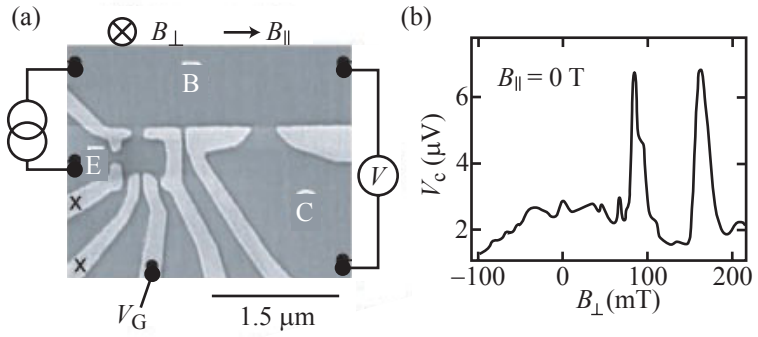


Fig. 13.9 (a) Polarizer–analyzer geometry with a quantum billiard as the emitter and a quantum point contact as the analyzer. (b) Magnetic focusing resonances at zero in-plane field (Folk *et al.*, 2003).

we find that an experiment with constant current bias is ideal if

$$G_{31} = \frac{h}{2e^2}\alpha\left(G_{11}^{\uparrow}G_{33}^{\uparrow} + G_{11}^{\downarrow}G_{33}^{\downarrow}\right),$$

i.e., if G_{31} can be split into the two spin contributions according to

$$G_{31}^{\uparrow/\downarrow} = \frac{h}{2e^2}\alpha G_{11}^{\uparrow/\downarrow}G_{33}^{\uparrow/\downarrow} \quad \text{with} \quad G_{31} = G_{31}^{\uparrow} + G_{31}^{\downarrow}.$$

This makes sense if the spin state of the electrons is conserved during their trip through the base B from the polarizer to the analyzer, implying negligible spin-flip scattering and spin–orbit interaction. The product $G_{11}^{\uparrow}G_{33}^{\uparrow}$, which is equivalent to the product of the corresponding transmission probabilities, excludes phase-coherent effects in the transmission from 1 to 3, and implies that the two transmission processes are statistically independent. In this case, the factor α can be interpreted as the probability that an electron emitted from the polarizer arrives at the analyzer.

The experiment shown in Fig. 13.8 shows that these conditions are fulfilled approximately as expressed by eq. (13.35). The heights of resonances in the focusing experiment depend at zero magnetic field only weakly on the opening of the two quantum point contacts. This opens the way for further experiments in which the polarizing quantum point contact is replaced by the quantum billiard structure. Such a setup is depicted in Fig. 13.9(a). Figure 13.9(b) shows the focusing resonances measured on this structure at zero in-plane field.

Figure 13.10 shows the corresponding measurement results for finite in-plane field. The billiard was coupled with a conductance of about $2e^2/h$ to the leads such that no spin polarization is expected from the point contact connecting the emitter to the base. The system is therefore open, and mesoscopic fluctuations of the conductance caused by interference of electronic waves inside the emitter billiard [Fig. 13.10(c)]. Significant fluctuations in the detected polarization signal do not arise if the emitter is a quantum point contact but only in the case of the confined emitter. If we compare the spin polarization signal with the conductance of the billiard [Fig. 13.10(c)], we find no correlation between the mesoscopic conductance fluctuations and the fluctuations of the spin polarization.

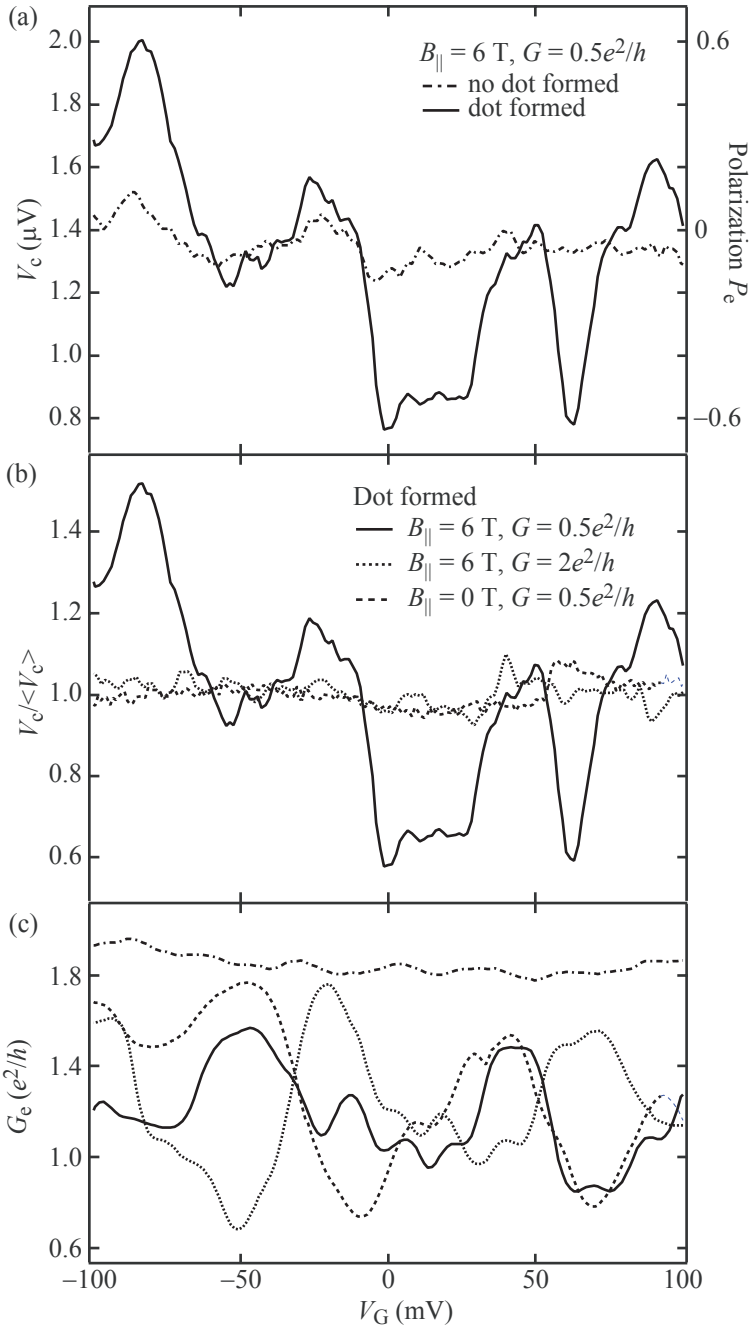


Fig. 13.10 (a) Measured collector voltage in the case where the emitter was operated as a quantum point contact with conductance $2e^2/h$ (solid) compared to the case where the quantum billiard emitter was formed (dash–dotted). Plotted is the height of a focusing resonance with an in-plane magnetic field of 6 T applied as a function of the plunger gate voltage of the emitter. (b) Normalized peak heights for the case of a spin-selective analyzer ($0.5e^2/h$, solid curve) compared to the non spin-selective analyzer (dotted) and for the case of zero in-plane magnetic field with an analyzer conductance of $0.5e^2/h$ (dashed) (c) The corresponding measurements of the emitter conductance (Folk *et al.*, 2003).

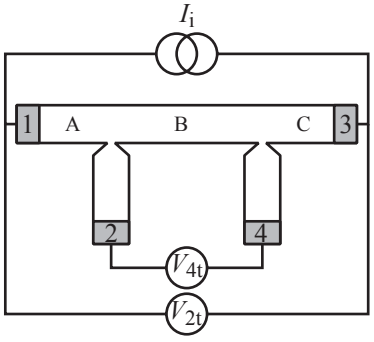


Fig. 13.11 Schematic of a mesoscopic Hall bar consisting of a wire with only a single mode.

Four-terminal resistance of a nearly perfect single-mode quantum wire. Consider a quantum wire with only a single mode measured in the four-terminal mesoscopic Hall bar arrangement depicted schematically in Fig. 13.11. Experimentally, such a wire with a length of more than 1 µm has been realized by the cleaved-edge overgrowth method. It was possible to measure the resistance of this wire in a four-terminal geometry. The philosophy behind the measurement of a macroscopic Hall bar is that the voltage drop across the inner two contacts 2 and 4 along the wire are related to the intrinsic resistance of the wire section B, while the voltage drop across the outer two contacts 1 (source) and 3 (drain) contains the contribution of the contact resistances (see Fig. 13.12). Now let us see how these ideas are modified in the mesoscopic system.

First we assume that the coupling strength of the two voltage leads to the wire is very weak. As a consequence, we will treat $\mathcal{T}_{13} = \mathcal{T}_{31}$ (zero magnetic field) as a large quantity of the order ϵ^0, whereas $\mathcal{T}_{12} = \mathcal{T}_{21}$, $\mathcal{T}_{14} = \mathcal{T}_{41}$, and $\mathcal{T}_{34} = \mathcal{T}_{43}$ are of order ϵ^1. The transmission $\mathcal{T}_{24} = \mathcal{T}_{42}$ is of the order ϵ^2. The injected current I_i is given by the external current source, and the measured voltages are expanded according to eq. (13.33).

Equation (13.19) gives, in the order ϵ^0, for the two-terminal voltage drop $V_{13}^{(0)}$ across the device,

$$V_{2t} = V_{13}^{(0)} = \frac{h}{e^2}\frac{1}{\mathcal{T}_{13}} I_i,$$

implying that the two-terminal resistance is

$$R_{2t}^{(0)} = R_{13,13}^{(0)} = \frac{h}{e^2}\frac{1}{\mathcal{T}_{13}},$$

as if the voltage probes were not attached to the wire.

In the experiment shown in Fig. 13.12 the two-terminal resistance shows pronounced plateaus in the resistance, resembling conductance quantization in the wire. The plateaus are not exactly at $h/e^2 n$ (n integer) because coupling from the two-dimensional electron regions into the wire, and ohmic contacts, add a series resistance.

In order to find a result for the four-terminal measurement, we need to investigate eqs (13.19)–(13.22) in the order ϵ^1 assuming $I_2 = 0$. From eq. (13.20), and using the identity $V_{14} = V_{34} + V_{13}$, we find

$$V_{34}^{(0)} = -\frac{\mathcal{T}_{14}^{(1)}}{\mathcal{T}_{14}^{(1)} + \mathcal{T}_{34}^{(1)}} V_{13}^{(0)}.$$

Similarly we find from eq. (13.22) with $V_{12} = V_{13} - V_{23}$

$$V_{23}^{(0)} = \frac{\mathcal{T}_{12}^{(1)}}{\mathcal{T}_{12}^{(1)} + \mathcal{T}_{23}^{(1)}} V_{13}^{(0)}.$$

In this order, the voltage difference $V_{24}^{(0)}$ is therefore

$$V_{4t} = V_{24}^{(0)} = V_{23}^{(0)} + V_{34}^{(0)} = \left(\frac{\mathcal{T}_{12}^{(1)}}{\mathcal{T}_{12}^{(1)} + \mathcal{T}_{23}^{(1)}} - \frac{\mathcal{T}_{14}^{(1)}}{\mathcal{T}_{14}^{(1)} + \mathcal{T}_{34}^{(1)}} \right) V_{13}^{(0)}.$$

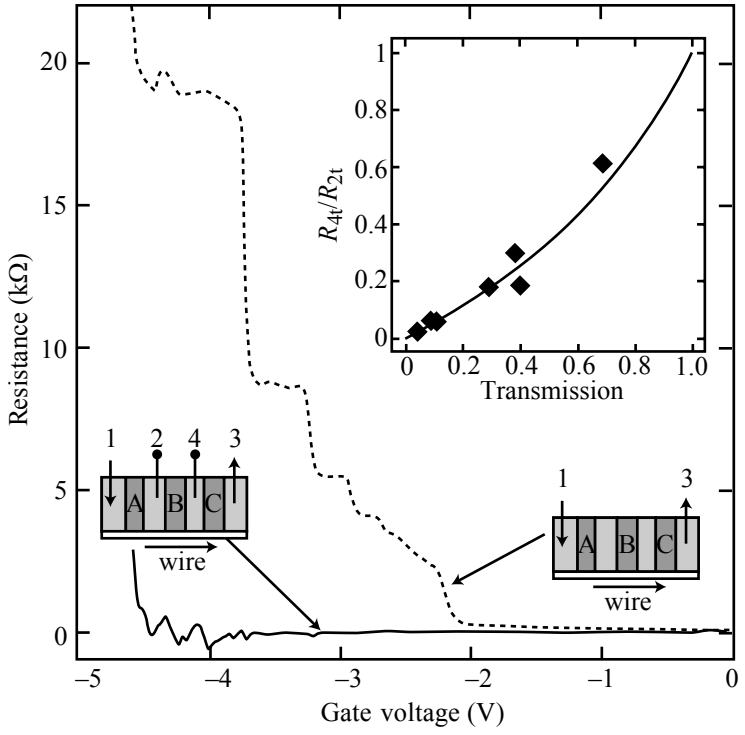


Fig. 13.12 Two- (dotted) and four-terminal (solid) resistance of a ballistic quantum wire fabricated by cleaved edge overgrowth. The dark gray regions are gate electrodes used to deplete the underlying two-dimensional electron gas. The light gray regions are the contacts to the wire. For the measurement, the varying gate voltage is applied to gate B. In the two-terminal measurement, gates A and C are not energized. (De Picciotto *et al.*, 2001. Reprinted by permission from Macmillan Publishers Ltd. Copyright 2001.)

It is interesting to see that in a mesoscopic Hall bar the voltage drop $V_2 - V_4$ is not necessarily positive, but it may take negative values, depending on how strongly the two voltage probes couple to the current contacts 1 and 3. This is a direct consequence of eq. (13.11) according to which the voltage on each terminal is a weighted average of the voltages on the other terminals transmitting to it. The corresponding four-terminal resistance is

$$R_{4t}^{(0)} = R_{13,24} = R_{2t}^{(0)} \left(\frac{\mathcal{T}_{12}^{(1)}}{\mathcal{T}_{12}^{(1)} + \mathcal{T}_{23}^{(1)}} - \frac{\mathcal{T}_{14}^{(1)}}{\mathcal{T}_{14}^{(1)} + \mathcal{T}_{34}^{(1)}} \right).$$

The two fractions in brackets give, in general, values between zero and one, such that their difference is bound to the interval $[-1; 1]$. The two expressions in the denominators can be identified with the emissivities $N_2 - \mathcal{R}_2 \approx \mathcal{T}_{12} + \mathcal{T}_{23}$, and $N_4 - \mathcal{R}_4 \approx \mathcal{T}_{14} + \mathcal{T}_{34}$ of the two voltage probes. For the experiments on quantum wires fabricated by cleaved edge overgrowth it was found that the emissivities are given by $2\alpha\mathcal{T}(W)$, where α is a constant and $\mathcal{T}(W)$ is a transmission probability depending on the width W of the probe. In the experiment discussed here, the width of the two voltage probes, and hence their emissivities, are the same.

So far we have not exploited eq. (13.19) in the order ϵ^1. It turns out that doing this gives a correction of the previous lowest order two-terminal resistance. This shows us that, in mesoscopic devices, attaching

voltage probes to a system will always tend to influence the resistance of the system, unlike the notion in macroscopic Hall bars, where the attachment of small voltage probes at the edge of the Hall bar does not have such an influence. To be specific, we find from eq. (13.19) in the order ϵ^1

$$V_{13}^{(1)} = -\frac{h}{e^2} I_{\mathrm{i}} \frac{1}{{\mathcal{T}_{13}^{(0)}}^2} \left(\frac{\mathcal{T}_{12}^{(1)} \mathcal{T}_{23}^{(1)}}{\mathcal{T}_{12}^{(1)} + \mathcal{T}_{23}^{(1)}} + \frac{\mathcal{T}_{14}^{(1)} \mathcal{T}_{34}^{(1)}}{\mathcal{T}_{14}^{(1)} + \mathcal{T}_{34}^{(1)}} \right)$$

such that

$$R_{2\mathrm{t}} = \frac{h}{e^2} \frac{1}{\mathcal{T}_{13}^{(0)}} \left[1 - \frac{1}{\mathcal{T}_{13}^{(0)}} \left(\frac{\mathcal{T}_{12}^{(1)} \mathcal{T}_{23}^{(1)}}{\mathcal{T}_{12}^{(1)} + \mathcal{T}_{23}^{(1)}} + \frac{\mathcal{T}_{14}^{(1)} \mathcal{T}_{34}^{(1)}}{\mathcal{T}_{14}^{(1)} + \mathcal{T}_{34}^{(1)}} \right) \right].$$

The two-terminal resistance in the order ϵ^1 is reduced compared to that in lowest order. The reason is that, in addition to the direct transmission $\mathcal{T}_{13}^{(0)}$, indirect transmission of electrons from contact 1 to 3 via contact 2, and from 1 to 3 via 4 are now also accounted for. This effectively enhances the transmission from contact 1 to 3, thereby reduces the voltage V_{13} (negative sign of $V_{13}^{(1)}$), and lowers the two-terminal resistance.

As mentioned above, in the experiment, the two voltage probes couple with identical strength $2\alpha\mathcal{T}(W)$ to the wire, because they are fabricated to have the same width W. The four-terminal resistance can then be written as

$$R_{4\mathrm{t}}^{(0)} = R_{13,24} = R_{2\mathrm{t}}^{(0)} \frac{\mathcal{T}_{12}^{(1)} - \mathcal{T}_{14}^{(1)}}{2\alpha\mathcal{T}(W)}.$$

The experiment shows (see Fig. 13.12) that the difference in the numerator is essentially zero for the gate voltage range between slightly below $-2\,\mathrm{V}$, where the two-dimensional electron gas in region B is depleted, and almost $-4\,\mathrm{V}$, where small resistance fluctuations set in. This difference can only be zero if the wire segment along gate B does not scatter electrons, i.e., if there is no intrinsic resistance. As a conclusion, the quantized resistance observed in Fig. 13.12 for the two-terminal measurement cannot be a result of scattering within the wire, but must be a result of the current contact regions, in agreement with our previous interpretation of the quantized conductance as a contact resistance.

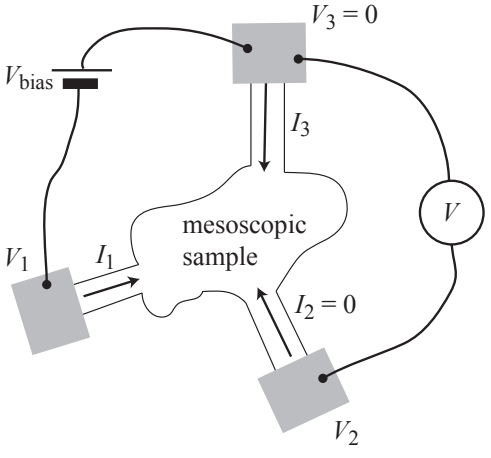
It turns out that the strength with which $\mathcal{T}_{13}$ is altered by the presence of the two voltage probes (the invasiveness of the probes) is proportional to $\mathcal{T}(W)$. The invasiveness decides whether the four-terminal measurement allows the measurement of the zero intrinsic wire resistance. The inset of Fig. 13.12 shows data points obtained from a series of devices differing in the contact width W, and therefore in the invasiveness $\mathcal{T}(W)$. Plotted is the ratio between the four- and the two-terminal resistance. It can be seen that only for the smallest values of $\mathcal{T}$, i.e., for the smallest widths W, does the four-terminal resistance measure the zero intrinsic wire resistance. As $\mathcal{T}$ approaches one, two- and four-terminal resistances become equal.

Further reading

- Landauer–Büttiker formalism: Datta 1997; Beenakker and van Houten 1991.
- 'Magnetic steering' and 'magnetic focusing': Beenakker and van Houten 1991.
- Papers: Buttiker 1986; Molenkamp *et al.* 1990; Potok *et al.* 2002; Potok *et al.* 2003; Folk *et al.* 2003; de Picciotto *et al.* 2001.

Exercises

(13.1) We investigate a three-terminal device which is based on a two-dimensional electron gas in the quantum limit. A schematic of the sample with external circuit is depicted below. We neglect the resistances of the ohmic contacts. The number of modes in the three ideal leads is N_i ($i = 1, 2, 3$), the reflection in contact i is $\mathcal{R}_i$, and the transmission from contact i to contact j is $\mathcal{T}_{ji}$.



(a) Calculate the relation between V/V_{bias} and the $\mathcal{T}_{ij}$ within Landauer–Büttiker theory.

(b) Discuss what it means for the transmissions $\mathcal{T}_{21}$ and $\mathcal{T}_{23}$ if $\mathcal{R}_2$ becomes very large. How large can $\mathcal{R}_2$ be at most?

(c) Discuss which voltage V is measured if $\mathcal{R}_2$ tends towards its maximum value. Experimentally this could, for example, be achieved by placing a gate across lead 2 and depleting the electron gas with a negative voltage.

(d) Show that, in this limit, the conductance of the structure approaches that of a two-terminal device.

This page intentionally left blank

Interference effects in nanostructures I

14

14.1 Double-slit interference

Interference of waves of matter is a cornerstone of quantum mechanics arising from the superposition principle. It plays an important role in all applications of quantum mechanical theory to real systems. For example, the existence of the band structure of a solid can be regarded as an effect of interference of partial electron waves multiply scattered at the ion cores of the crystal lattice. However, the double slit experiment is probably one of the oldest and most striking ways to discuss and discover interference and the wave–particle duality of matter.

Figure 14.1 shows schematically the setup of a double slit experiment. A coherent monochromatic source of waves of light or matter creates an outgoing wave that is diffracted by two slits with a width smaller than the wavelength. Under these conditions an interference pattern is observed on the observation screen, i.e., the intensity of the impinging wave varies periodically in space.

A large number of such experiments have been realized in the past in very different areas of physics. Among them are those using classical waves such as sound or water waves. Among them is the famous double slit experiment with light, i.e., photons, that Thomas Young performed in 1801. At that time the double slit experiment was a key experiment

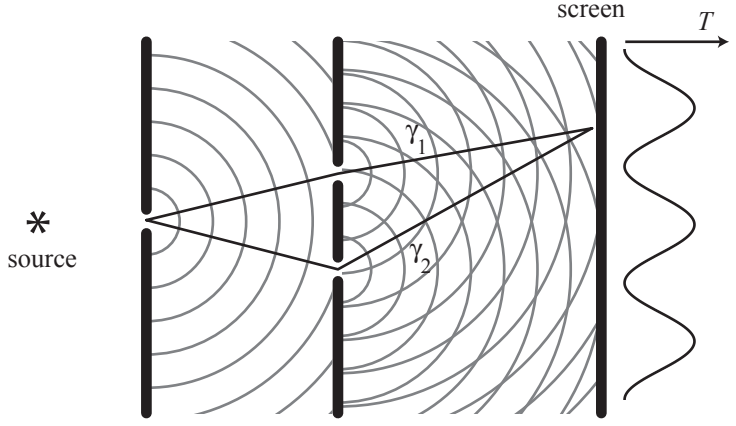


Fig. 14.1 Schematic setup of a double slit experiment.

interpreted as a proof for the wave nature of light. With today's knowledge of quantum mechanics, of course, we interpret this experiment in a different way: even when the intensity of the incident photon beam is reduced so much that only a single photon is present at a time in the apparatus, the interference pattern is formed, if one counts the number of photons arriving at a particular place on the screen. Here, the important step is made from a (real valued) amplitude of a classical wave to the (complex) probability amplitude of a quantum mechanical particle. The probability amplitude t_1 for transmission through the upper slit along path γ_1 in Fig. 14.1 is described as

$$t_1 = a_1 e^{i\theta_1}.$$

Correspondingly, the transmission t_2 along path γ_2 is

$$t_2 = a_2 e^{i\theta_2}.$$

Here, a_1 and a_2 are positive real numbers between zero and one, the θ_i $(i = 1, 2)$ are real-valued transmission phases. According to the rules of quantum mechanics, the intensity on the observation screen is given by

$$\mathcal{T} = |t_1 + t_2|^2 = a_1^2 + a_2^2 + 2a_1 a_2 \cos\delta, \text{ with } \delta = \theta_1 - \theta_2. \qquad (14.1)$$

The first two terms of the transmission $a_1^2 + a_2^2$ can be interpreted as the classical transmission probability (i.e., the sum of the two individual probabilities for transmission through either of the two slits), the last term $2a_1 a_2 \cos\delta$ represents the quantum interference.

In the 20th century the double slit experiment was performed with a large number of different particles, such as electrons, neutrons, atoms, and even molecules [C_{60}, so-called *bucky balls* (Arndt *et al.*, 1999)]. The first man-designed double slit type of experiments with electrons were performed by Möllenstedt and Düker, and by Jönsson with the electron beam of a scanning electron microscope (Möllenstedt and Duker, 1956; Jönsson, 1961). All these experiments show the expected interference pattern. However, interference experiments remain of interest in research. The reason is that fundamental phenomena of phase-coherence and decoherence of quantum particles, and its relation to *entanglement* of quantum states can be studied. Furthermore, these phenomena are of crucial importance for possible applications, such as *quantum information processing* or *quantum communication*.

14.2 The Aharonov–Bohm phase

If one intends to realize a double slit experiment in a semiconductor nanostructure, a number of problems arise: in nanostructures, particles are detected in contacts, i.e., it is difficult to realize an interference screen. As a way out, the Aharonov–Bohm effect (Aharonov and Bohm, 1959, see also Ehrenberg and Siday, 1949, for an early version of the effect) depicted schematically in Fig. 14.2 can be used. For this effect to

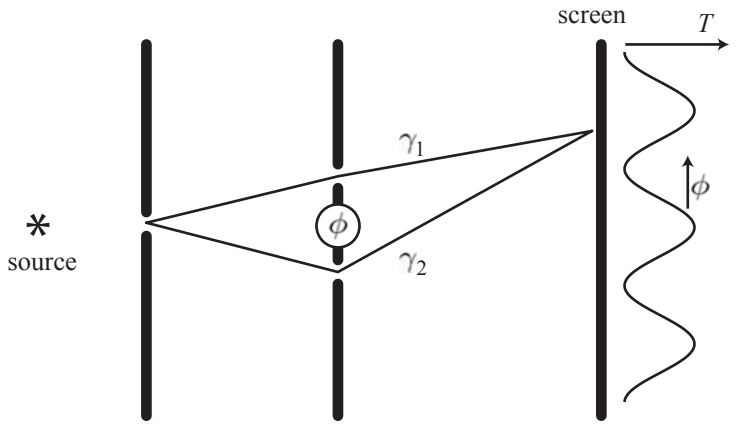


Fig. 14.2 Schematic setup of an Aharonov–Bohm experiment.

occur, the two interfering paths have to enclose a magnetic flux ϕ that can be described by a vector potential $\mathbf{A}(\mathbf{r})$. The transmission phases are then modified by the presence of the magnetic flux according to

$$\theta_i(\phi) = \theta_i(0) - \frac{|e|}{\hbar}\int_{\gamma_i} \mathbf{A} d\mathbf{s}. \tag{14.2}$$

As a result, the phase difference,

$$\delta(\phi) = \delta(0) - \frac{|e|}{\hbar}\int_{\gamma_1-\gamma_2} \mathbf{A} d\mathbf{s} = \delta(0) - 2\pi\frac{\phi}{\phi_0}, \tag{14.3}$$

depends on the magnetic flux ϕ, where $\phi_0 = h/|e|$ is the magnetic flux quantum. As a consequence, the flux-dependent transmission is

$$\mathcal{T}(\phi) = a_1^2 + a_2^2 + 2a_1a_2\cos\left[\delta(0) - 2\pi\frac{\phi}{\phi_0}\right].$$

At a fixed position on the interference screen, the transmission has the property

$$\mathcal{T}(\phi + n\cdot\phi_0) = \mathcal{T}(\phi), \text{ with } n \text{ integer}.$$

In this way it is possible to detect the periodic interference pattern with a spatially fixed detector by changing the magnetic flux. The additional phase $2\pi\phi/\phi_0$ appearing in the transmission probability is called the *Aharonov–Bohm phase*.

Circular motion. The above considerations were quite general. No constraints were put on the particular geometric shape of the two paths $\gamma_{1,2}$. Now we illustrate the appearance of the Aharonov–Bohm phase using the specific example of an electron moving on a circular path around a magnetic flux, as depicted in Fig. 14.3(a). The moving charge $-|e|$ can be described using the hamiltonian

$$H = \frac{1}{2m}\left(\mathbf{p} + |e|\mathbf{A}\right)^2 + V(r),$$

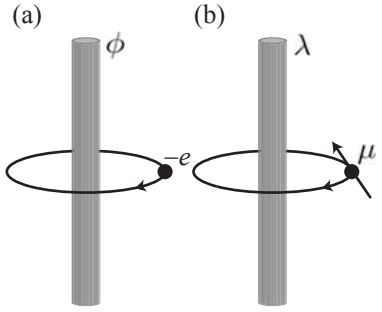


Fig. 14.3 Schematic illustration of the settings for the Aharonov–Bohm effect and its electromagnetic dual, the Aharonov–Casher effect. (a) In case of the Aharonov–Bohm effect a charged particle (e.g., charge $-|e|$) is encircling a magnetic flux tube enclosing the flux ϕ. (b) In case of the Aharonov–Casher effect, an uncharged particle with a magnetic moment (spin) is encircling a tube of constant line charge density λ.

where the vector potential $\mathbf{A} = \mathbf{e}_\varphi \phi/2\pi r$ ($\mathbf{e}_\varphi$ is a unit vector pointing normal to the radial direction in the plane) and r measures the distance from the flux tube. The potential $V(r)$ forces the electron into the circular orbit, motion in the z-direction is assumed to be separable. Owing to the fact that the charge moves in a region free of magnetic field (because $\mathbf{B} = \nabla \times \mathbf{A} = 0$ for $r > 0$) it does not experience a Lorentz force. Nevertheless, as the quantum particle moves once around the flux tube it acquires the Aharonov–Bohm phase

$$\Delta\varphi_{\mathrm{AB}} = -\frac{|e|}{\hbar} \oint \mathbf{A} d\mathbf{s} = -\frac{|e|}{\hbar}\phi = -2\pi\frac{\phi}{\phi_0}. \tag{14.4}$$

This can be seen by writing the above hamiltonian in cylinder coordinates

$$H = -\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} - \frac{1}{r^2}\left(i\frac{\partial}{\partial\varphi} - \frac{\phi}{\phi_0}\right)^2\right] + V(r).$$

The one-dimensional hamiltonian for pure angular motion can be obtained by neglecting the radial derivative terms and letting $r = r_0$, leading to

$$H = \frac{\hbar^2}{2mr_0^2}\left(i\frac{\partial}{\partial\varphi} - \frac{\phi}{\phi_0}\right)^2.$$

The eigenvalue problem is solved by the wave function

$$\psi(\varphi) = \frac{1}{\sqrt{2\pi}} e^{i\ell\varphi}$$

with the angular momentum quantum number ℓ, and having the eigenenergy

$$E_\ell = \frac{\hbar^2}{2mr_0^2}\left(\ell + \frac{\phi}{\phi_0}\right)^2. \tag{14.5}$$

This dispersion relation is parabolic as depicted in Fig. 14.4. The sign $\lambda = \mathrm{sgn}(v_{\mathrm{G}})$ of the group velocity

$$v_{\mathrm{G}}(\ell) = \frac{r_0}{\hbar}\frac{\partial E_\ell}{\partial \ell} = \frac{\hbar}{mr_0}\left(\ell + \frac{\phi}{\phi_0}\right) \tag{14.6}$$

indicates the direction of propagation around the ring. For each energy E of an electron, two states propagating in opposite directions exist, because according to eq. (14.5)

$$\ell + \frac{\phi}{\phi_0} = \lambda\sqrt{\frac{2mr_0^2 E}{\hbar^2}}$$

with $\lambda = \pm 1$ indicating the direction of propagation.

In the case of an isolated ring [as depicted in Fig. 14.3(a)], the eigenfunctions need to be periodic in φ leading to integer angular momentum quantum numbers ℓ. Positive $\ell + \phi/\phi_0$ describe states propagating counterclockwise around the ring, for negative $\ell + \phi/\phi_0$ states propagate clockwise.

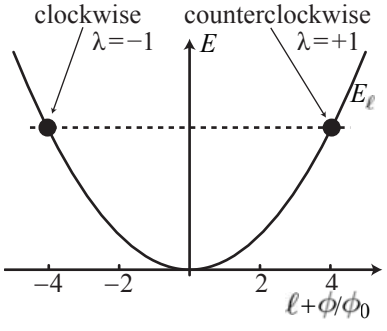


Fig. 14.4 Dispersion relation for the one-dimensional ring threaded by a magnetic flux. The two states marked with a filled circle have the same energy, but differ in the direction λ of propagation around the ring.

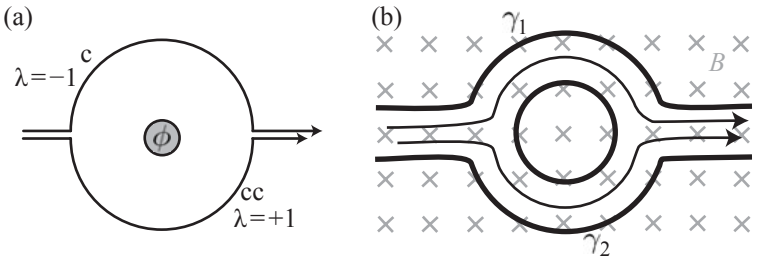


Fig. 14.5 (a) Schematic illustration of counterclockwise ('cc') and clockwise ('c') transmission around an open ring. Paths with higher winding numbers, or reflections at the joints to the contacts are neglected. (b) Quantum ring geometry used for realizing an Aharonov–Bohm experiment in a semiconductor nanostructure. A homogeneous magnetic field B is applied normal to the plane of the electron gas.

If the ring is open to leads attached at $\varphi = 0$ and $\varphi = \pi$, ℓ is not required to be integer because we are dealing with an open system. States impinging from one lead onto the ring with a particular energy E will be transmitted through the ring in one of the two states existing at this energy. Defining the wave vector $k = \sqrt{2mE/\hbar^2}$, we can solve the energy dispersion for ℓ and find the two states

$$\ell^{(\lambda)} = \lambda k r_0 - \frac{\phi}{\phi_0}.$$

These two states propagate with the same magnitude $\hbar k/m$ of their group velocity in counterclockwise ($\lambda = 1$) and clockwise ($\lambda = -1$) directions, respectively, as illustrated in Fig. 14.5(a). We now consider an incoming electron that has been split by the beam splitter at the ring entrance into two partial waves each having an amplitude $1/\sqrt{2}$ and propagating clockwise and counterclockwise, respectively. The two partial waves meet at the exit, exactly opposite to the entrance. The wave propagating counterclockwise travels an angle π from the entrance to the exit and acquires a phase factor $t_1 = \exp[i(kr_0 - \phi/\phi_0)\pi]/\sqrt{2}$. The wave propagating clockwise travels an angle $-\pi$ from the entrance to the exit and acquires a phase factor $t_2 = \exp[i(-kr_0 - \phi/\phi_0)(-\pi)]/\sqrt{2}$. The probability for the electron to be transmitted from the entrance point to the exit point within half a revolution around the ring is then given by

$$\mathcal{T} = |t_1 + t_2|^2 = \frac{1}{2}\left|e^{i(kr_0 - \phi/\phi_0)\pi} + e^{i(kr_0 + \phi/\phi_0)\pi}\right|^2 = \frac{1}{2}[1 + \cos(2\pi\phi/\phi_0)].$$

The transmission is modulated by the Aharonov–Bohm phase (14.4). In this simplified picture, we have neglected any reflection at the entrance to the ring and assumed that the particles arriving at the exit are perfectly coupled out.

14.3 Aharonov–Bohm experiments

In principle, an Aharonov–Bohm experiment like that depicted in Fig. 14.2 can be realized in a semiconductor nanostructure based on a two-dimensional electron gas by using split gates. However, the edges of

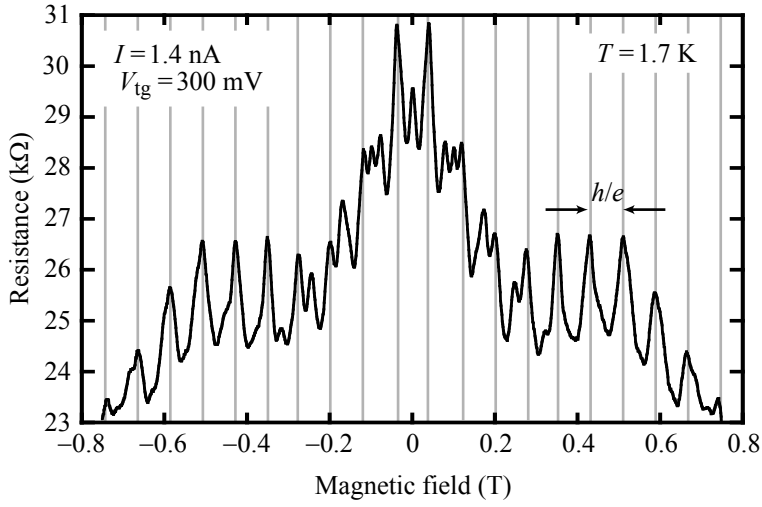


Fig. 14.6 Aharonov–Bohm effect in the quantum ring structure shown in Fig. 6.15(a).

the sample deserve particular attention. In the ideal experiment, partial waves that are scattered between the source and the double slit, or between the double slit and the detector, can escape the structure sideways. In a nanostructure realization, these regions would therefore need to be connected to ground via ohmic contacts. Doing this, a lot of intensity of the interference pattern is lost. Therefore, more closed structures such as quantum rings [see Fig. 14.5(b)] have been used for measuring the Aharonov–Bohm effect. Here, another important difference from the ideal experiment in Fig. 14.2 is the application of a homogeneous magnetic field. As a result, the electrons experience a Lorentz force in addition to acquiring the Aharonov–Bohm phase. The Lorentz force leads to cyclotron motion of the free electron in the magnetic field. In the structure depicted in Fig. 14.2, however, the electrons are not free, but bound to the ring-shaped structure. As a consequence, there is an interplay between forces caused by the confinement potential, and the Lorentz force caused by the magnetic field. As a rule of thumb we can say that Lorentz force effects are negligible as long as the classical cyclotron radius of the electron $R_c = p/eB = \hbar k_F/eB$ is large compared to the radius of the ring structure. This is given for sufficiently small magnetic fields. In this magnetic field range we expect that the transmission through the structure is mainly modulated by the Aharonov–Bohm effect for which the relevant magnetic flux is given by

$$\phi = B \cdot A.$$

Here, A is the mean area enclosed by the ring structure. Figure 6.15 shows such a structure that has been fabricated by AFM lithography.

The corresponding measurement of the magnetoconductance at a temperature of 1.7 K is shown in Fig. 14.6. It shows a very clear periodic modulation of the resistance as a function of magnetic field. The most prominent period ΔB corresponds to the expected h/e-periodic trans-

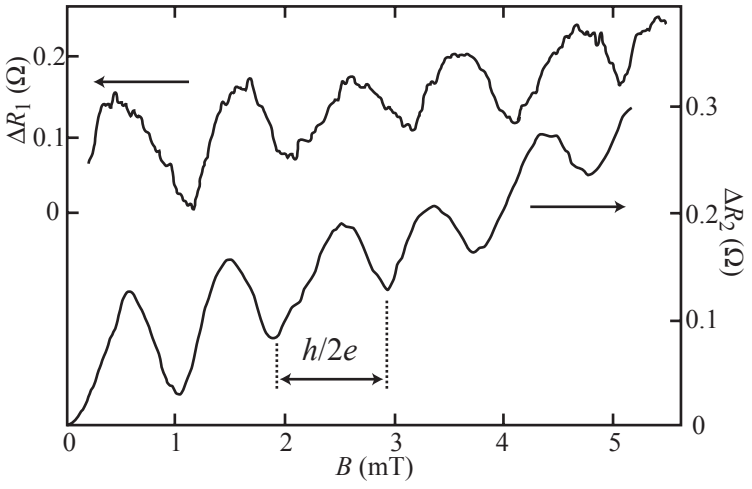


Fig. 14.7 Altshuler–Aronov–Spivak oscillations in a metal cylinder. (Reprinted with permission from Sharvin and Sharvin, 1981. Copyright 1981, American Institute of Physics.)

mission with

$$\Delta B = \frac{h/e}{A}.$$

In addition, higher harmonics can be seen, e.g., $h/2e$-periodic oscillations with

$$\Delta B = \frac{h/(2e)}{A}.$$

This is a characteristic property of quantum rings with high quality and weak decoherence.

Sharvin–Sharvin experiment. The Aharonov–Bohm effect was, for the first time in a solid structure, observed by Sharvin and Sharvin in a long magnesium cylinder evaporated on a micrometer-thin quartz filament (Sharvin and Sharvin, 1981). However, in this structure, only $h/(2e)$-periodic oscillations were seen (see Fig. 14.7) as predicted theoretically by Altshuler, Aronov, and Spivak (Altshuler *et al.*, 1981). These $h/2e$-periodic oscillations are therefore often called Altshuler–Aronov–Spivak (AAS) oscillations (in contrast to the h/e-periodic Aharonov–Bohm (AB) oscillations).

h/e-periodic oscillations in metals and semiconductors. The h/e-periodic Aharonov–Bohm oscillations were first measured in a single metal ring by R. Webb and co-workers (Webb *et al.*, 1985). The first measurements in rings based on semiconductor heterostructures were performed by Timp *et al.*, 1987, Ishibashi *et al.*, 1987, and Ford *et al.*, 1988.

Origin of higher harmonics. In order to understand the observation of higher harmonics better, we consider a one-dimensional model and

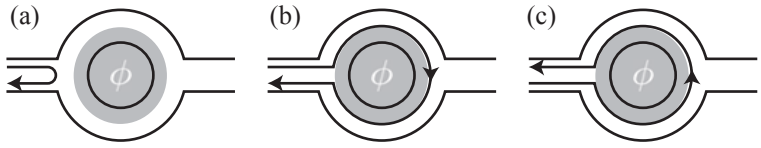


Fig. 14.8 Paths considered for the description of the Aharonov–Bohm experiment in a quantum ring structure. (a) The electron is reflected at the entrance to the ring. (b) The electron is reflected after having explored the ring once in a clockwise direction. (c) The electron is reflected after having explored the ring once in a counterclockwise direction.

use the Landauer–Büttiker formalism. We write the conductance at temperature zero as

$$G = \frac{2e^2}{h}\mathcal{T}(E_\mathrm{F}) = \frac{2e^2}{h}\left[1 - \mathcal{R}(E_\mathrm{F})\right]. \tag{14.7}$$

Here, $\mathcal{R}$ is the reflection probability for an electron at the Fermi energy. If we assume strong coupling of the ring to the leads, i.e., $\mathcal{R}(E_F) \ll 1$, we can limit our considerations to the paths depicted schematically in Fig. 14.8. We therefore obtain

$$\begin{aligned}
\mathcal{R} &= \left| r_0 + r_1 e^{i\cdot 2\pi\phi/\phi_0} + r_1 e^{-i\cdot 2\pi\phi/\phi_0} + \ldots \right|^2 \\
&= |r_0|^2 + 2|r_1|^2 + \ldots && (14.8)\\
&\quad +4\,|r_0|\,|r_1|\cos\delta\cos\left(2\pi\frac{\phi}{\phi_0}\right) + \ldots && (14.9)\\
&\quad +2\,|r_1|^2\cos\left(4\pi\frac{\phi}{\phi_0}\right) + \ldots && (14.10)
\end{aligned}$$

Here, δ is the phase that a partial wave accumulates if it runs once around the ring at zero magnetic field ($\phi = 0$). The contributions to the reflection can be classified in the following way:

- The first contribution (14.8) is independent of the magnetic flux ϕ and therefore also of the magnetic field B. It can be interpreted as the sum of classical reflection probabilities. We define $\mathcal{R}_\mathrm{cl} = |r_0|^2 + 2|r_1|^2$. The magnitude of these reflection probabilities does depend on the details of the local electrostatic potential within the ring and the coupling to the leads.
- The second contribution (14.9) is h/e-periodic and represents the Aharonov–Bohm effect. In this model, the Aharonov–Bohm effect appears as the interference between a path that is directly reflected at the entrance to the ring [Fig. 14.8(a)], and a path winding once around the ring [Fig. 14.8(b) or (c)]. The prefactor $\cos\delta$ can take arbitrary values between -1 and $+1$. This means that at $B = 0$ there can be a minimum or a maximum of the oscillations depending on the sign of this prefactor. The prefactor may also depend on the energy of the electron, because δ depends on energy. This has important consequences: if we calculate the reflection probability

at finite temperatures, we essentially average the energy-dependent reflection over an energy interval of about $4k_BT$. If $\delta(E)$ changes by more than π in this interval, the averaging procedure results in a reduction of the oscillation amplitude. Similar arguments are valid for ensemble averages. If one measured a parallel, or a serial connection of many Aharonov–Bohm rings, each ring would have a different value of the phase δ, and the amplitude of Aharonov–Bohm oscillations is strongly damped by averaging. The same argument is true beyond the model of a strictly one-dimensional ring if several radial modes exist in the ring (corresponding to several paths probing the local potential in different locations). Each of the modes (or paths) contributes with a different δ, and a reduction of the Aharonov–Bohm oscillation amplitude results.

- The third contribution (14.10) is $h/2e$-periodic. These are the Altshuler–Aronov–Spivak oscillations. They result from the interference of so-called time-reversed paths propagating clockwise and counterclockwise, respectively [Fig. 14.8(b) and (c)]. They can also be seen in the measurements shown in Fig. 14.6. At $B = 0$ ($\phi = 0$) these paths always give a positive contribution to the reflection, i.e., the partial waves interfere constructively in the reflection and thereby always lower the conductance at $B = 0$. This contribution is independent of the phase δ. Therefore, AAS oscillations are more robust against averaging than the h/e-periodic AB oscillations. Energy averaging, ensemble averaging, and averaging due to different radial modes (paths) around the ring do not lead to such a strong reduction of the AAS oscillation amplitude as seen for AB oscillations.

All contributions to the reflection have the common property that they are even in magnetic field, such that $G(B) = G(-B)$. This property is known as *phase-rigidity*, meaning that the phase of the AB oscillations can be 0 or π, that of the AAS oscillations is always 0. It is characteristic for all two-terminal experiments, i.e., for samples with only two contacts. The phase of the AB oscillations can change as a function of an experimental parameter, e.g., a gate voltage, or the electron density, if $\cos\delta$ goes through zero. The phase of the oscillations will then jump by π at the point where $\cos\delta = 0$. This is shown in Fig. 14.9. It can be seen that the amplitude of the AB oscillations diminishes to zero due to a zero of $\cos\delta$, but beyond it recovers with the sign of the oscillations reversed. In contrast, the AAS oscillations always show a maximum at $B = 0$.

Limits of this description. In the above model we have neglected paths winding around the ring more than once (we talk about paths with higher winding number). They would lead to further higher harmonics with a period $h/(n \cdot e)$ (n integer). Nevertheless many observations can be interpreted qualitatively on the basis of our strongly simplified model. The limits of this description become obvious if we consider the values

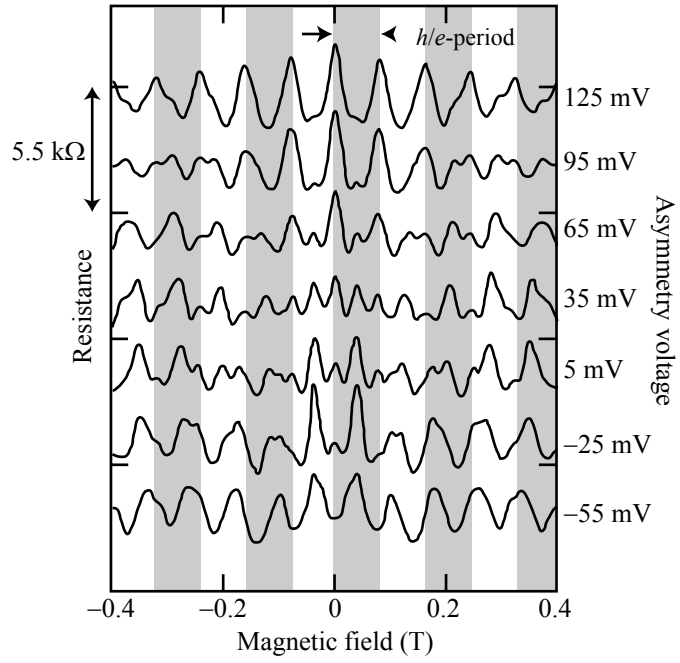


Fig. 14.9 Change of the phase of Aharonov–Bohm oscillations as a function of an asymmetrically applied gate voltage. In the topmost curve a maximum can be seen at $B = 0$, whereas a minimum is observed at $B = 0$ in the bottom curve. At the transition around 35 mV the amplitude of the h/e-periodic oscillations is zero and only the $h/2e$-periodic Altshuler–Aronov–Spivak oscillations can be seen.

of the transmission for different parameters $|r_0|$ and $|r_1|$. To this end we express $|r_0|$ and $|r_1|$ by $\mathcal{R}_{\rm cl}$ and $\Delta\mathcal{R}_{\rm cl} = |r_0|^2 - 2|r_1|^2$, and we write the transmission as

$$\mathcal{T} = 1 - \mathcal{R}_{\rm cl} - \sqrt{2(\mathcal{R}_{\rm cl}^2 - \Delta\mathcal{R}_{\rm cl}^2)}\cos\delta\cos\left(2\pi\frac{\phi}{\phi_0}\right) - \frac{1}{2}(\mathcal{R}_{\rm cl} - \Delta\mathcal{R}_{\rm cl})\cos\left(4\pi\frac{\phi}{\phi_0}\right) - \dots$$

The Aharonov–Bohm oscillations are strongest for the case $\cos\delta = \pm 1$ and $\Delta\mathcal{R}_{\rm cl} = 0$. For these extreme parameters we obtain the minimum for the transmission ($\phi/\phi_0 = 0$)

$$\mathcal{T}_{\rm min} = 1 - \left(\frac{3}{2} + \sqrt{2}\right)\mathcal{R}_{\rm cl} - \dots$$

Due to the constraint $\mathcal{T}_{\rm min} \geq 0$ our approximation delivers no physically meaningful results for $\mathcal{R}_{\rm cl} > 0.34$. In this case, higher harmonics are required for the description of the transmission.

Temperature dependence. At room temperature Aharonov–Bohm oscillations are usually not observable. Figure 14.10 shows measurements of the magnetoresistance of the quantum ring at different temperatures. Each type of quantum oscillation smooths out at increasing temperatures. This corresponds to our everyday experience that quantum interference effects can usually not be observed at room temperature. But why do these oscillations disappear with increasing temperature?

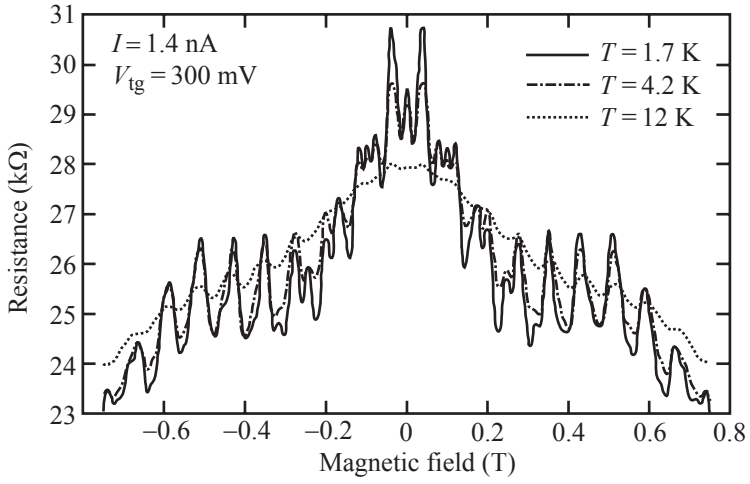


Fig. 14.10 Temperature dependence of the magnetoresistance of a quantum ring.

One influence of temperature on the magnetoconductance is contained in the derivative of the Fermi distribution function that appears in eq. (13.9) under the energy integral [so far in this chapter we have treated the conductance at temperature $T = 0$, cf., eq. (14.7)]. Since the transmission $\mathcal{T}$ (and the reflection $\mathcal{R}$) are functions of the energy, the conductance represents an average of the transmission over energy intervals of size $4k_BT$. The h/e-periodic AB oscillations in (14.9) contain the energy-dependent prefactor $\cos\delta$ which can be positive or negative depending on energy. Thermal averaging over large energy intervals therefore leads to vanishing AB oscillations.

For the $h/(2e)$-periodic AAS oscillations in (14.10) the situation is different. There, the prefactor is always positive and it usually depends only weakly on energy. This contribution to the oscillatory magnetoresistance is therefore less sensitive to an increase in temperature.

The temperature dependence of the amplitudes of AB and AAS oscillations can be extracted from measurements by Fourier analysis. The result of such a procedure is shown in Fig. 14.11. One can see that the $h/2e$-periodic oscillations decay more rapidly than the h/e-periodic oscillations. The reason for this behavior is the decoherence of electron waves upon traversal of the ring. This effect will be discussed in more detail in section 14.7.

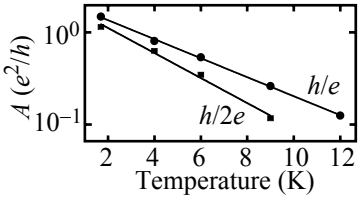


Fig. 14.11 Temperature dependence of the amplitude A of h/e- and $h/2e$-periodic oscillations as determined from a Fourier analysis of the data in Fig. 14.10.

14.4 Berry's phase and the adiabatic limit

Magnetic fields in ring structures can give rise to effects in the interference of partial waves beyond the Aharonov–Bohm effect if the spin of the charge carriers is taken into account, for example, through the Zeeman interaction. It has been proposed that these effects appear if the magnetic field not only has a component normal to the plane, but also a radial or tangential component in the plane of the ring (Loss *et al.*, 1990;

Loss and Goldbart, 1992; Stern, 1992). Such a situation has been called a ring in a *textured* magnetic field. As we will see below, the orbital motion around the ring causes the local spin orientation to deviate from the local direction of the magnetic field, in stark contrast to the case of a spin at rest which is oriented either parallel or antiparallel to the external field. The latter situation can be recovered in the *adiabatic limit* in which the Zeeman interaction dominates over the orbital motion. The adiabatic limit is of great interest because the phase acquired by the particle in an interference experiment has a geometric meaning, as we will see below. It is therefore called *geometric phase*, or *Berry's phase* (Berry, 1984). Geometric phases are interesting, because they arise from fundamental quantum mechanical principles and occur in many different physics contexts (Berry, 1988; Shapere and Wilczek, 1989). The concept of geometric phases was known before Berry's seminal paper [see, e.g., Schiff, 1949], but Berry realized its relevance for adiabatic dynamics in quantum mechanics. Experimentally, effects of Berry's phase have been demonstrated, for example, for photons (Tomita and Chiao, 1986), and for neutrons (Bitter and Dubbers, 1987).

In order to discuss an example where Berry's phase is expected to occur in an interference experiment using a quantum ring structure, we consider an electron with mass m in a strictly one-dimensional ring of radius r_0 described by the hamiltonian

$$H = \frac{\hbar^2}{2mr_0^2}\left(-i\frac{\partial}{\partial\varphi} + \frac{\phi}{\phi_0}\right)^2 + \frac{1}{2}g\mu_{\mathrm{B}}\mathbf{B}\boldsymbol{\sigma}.$$

The first term describes the kinetic energy with the magnetic flux $\phi = B_z\pi r_0^2$. The second term is the Zeeman interaction with the g-factor g and the magnetic field $\mathbf{B}$. We choose it to be given by

$$\mathbf{B} = B_z\mathbf{e}_z + B_r\mathbf{e}_r,$$

where we use a cylindrical coordinate system with $\mathbf{e}_r$ being a unit vector pointing radially outwards and $\mathbf{e}_z$ a unit vector in the z-direction. This situation is shown in Fig. 14.12. The magnetic field $\mathbf{B}$ has the magnitude $B = \sqrt{B_r^2 + B_z^2}$, and it is at an angle α to the z-axis with $\tan\alpha = B_r/B_z$. We note here that as the electron moves around the ring it experiences a magnetic field of varying orientation.

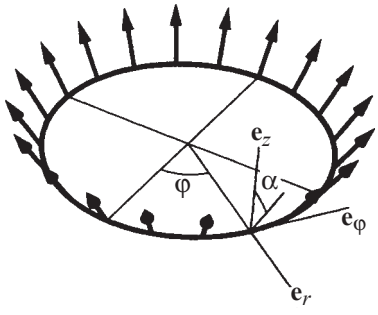


Fig. 14.12 Ring (thick circle in a magnetic field texture (arrows) with tilt angle α. The local cylindrical coordinate system at φ is indicated by $\{\mathbf{e}_r, \mathbf{e}_\varphi, \mathbf{e}_z\}$. (Reprinted with permission from Loss *et al.*, 1990. Copyright 1990 by the American Physical Society.)

The one-dimensional eigenspinors that solve this problem have the general form

$$\psi(\varphi) = \frac{1}{\sqrt{2\pi}}e^{i\ell\varphi}\begin{pmatrix}\chi_1\\ \chi_2 e^{i\varphi}\end{pmatrix}. \tag{14.11}$$

Inserting this expression in the eigenvalue problem we find the matrix equation

$$\begin{pmatrix}(\lambda-\frac{1}{2})^2 + \beta B_z - \epsilon & \beta B_r\\ \beta B_r & (\lambda+\frac{1}{2})^2 - \beta B_z - \epsilon\end{pmatrix}\begin{pmatrix}\chi_1\\ \chi_2\end{pmatrix} = 0 \tag{14.12}$$

for the two spin components, where $\beta = g\mu_{\mathrm{B}}mr_0^2/\hbar^2$ and $\lambda = \ell + \phi/\phi_0 + 1/2$. Before we look at the solutions of this eigenvalue problem we have

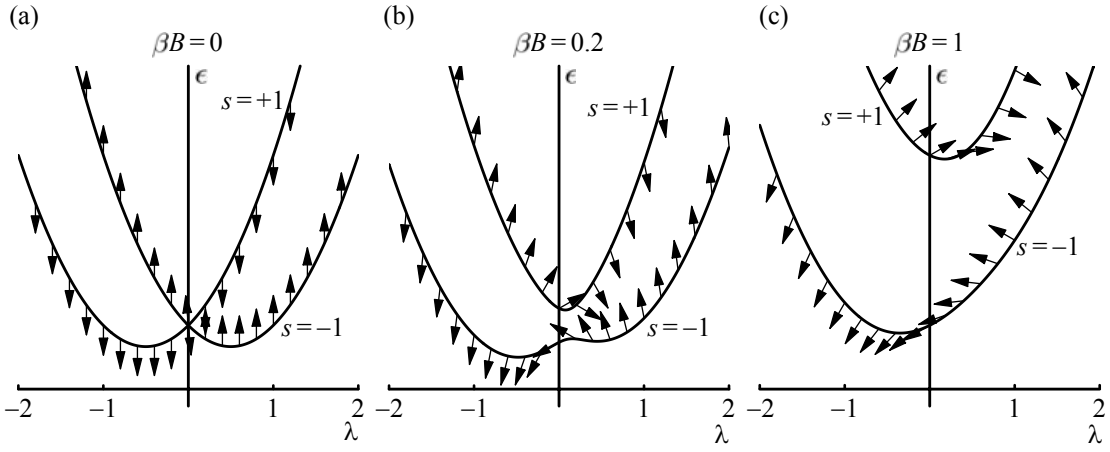


Fig. 14.13 Dispersion relations for electrons in a one-dimensional ring subject to a textured magnetic field. The tilt angle α of the field was chosen to be $\pi/3$, the spin orientation at $\varphi = 0$ for states along the dispersion is indicated by arrows. (a) The case of zero magnetic field. (b) An avoided crossing appears at finite magnetic fields. (c) The Zeeman splitting dominates the dispersion.

to realize that it is identical to the Zeeman problem of a spin (at rest) in a magnetic field except for the additional kinetic energy terms $(\lambda \mp 1/2)^2$. We therefore expand the spinors (χ_1, χ_2) in eigenspinors of the Zeeman problem, i.e.,

$$\begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} = c_1 \begin{pmatrix} \cos\frac{\alpha}{2} \\ \sin\frac{\alpha}{2} \end{pmatrix} + c_2 \begin{pmatrix} -\sin\frac{\alpha}{2} \\ \cos\frac{\alpha}{2} \end{pmatrix}. \tag{14.13}$$

The eigenspinors of the Zeeman problem are always parallel or antiparallel to the direction of the magnetic field. The transformed eigenvalue problem is then

$$\begin{pmatrix} \lambda^2 + \frac{1}{4} + \beta B - \lambda\cos\alpha - \epsilon & \lambda\sin\alpha \\ \lambda\sin\alpha & \lambda^2 + \frac{1}{4} - \beta B + \lambda\cos\alpha - \epsilon \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = 0. \tag{14.14}$$

The exact eigenenergies are found from eq. (14.14) to be

$$\epsilon_{\lambda,s} = \lambda^2 + \frac{1}{4} + s\sqrt{\beta^2 B_r^2 + (\lambda - \beta B_z)^2}. \tag{14.15}$$

Dispersion relations for different parameters βB are depicted in Fig. 14.13. In the case (a) of zero magnetic field, the dispersion consists of two parabolae displaced by one in the horizontal direction. At finite fields (b), an avoided crossing appears around $\lambda = 0$ which is of the order of the Zeeman splitting. At even higher fields (c), the Zeeman splitting starts to dominate the dispersion.

We write the eigenspinors (c_1, c_2) as

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \cos\frac{\delta}{2} \\ \sin\frac{\delta}{2} \end{pmatrix}, \tag{14.16}$$

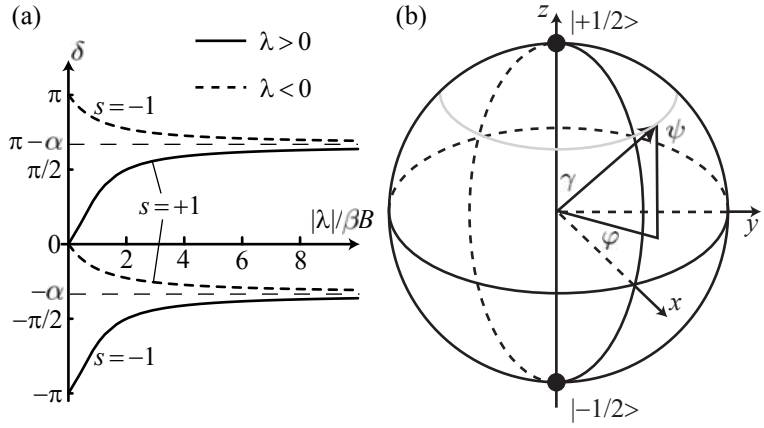


Fig. 14.14 (a) Tilt angle δ quantifying the deviation of the spin orientation from the direction of the magnetic field. The angle was calculated for an orientation $\alpha = \pi/3$ of the magnetic field texture. (b) Representation of $\psi(\varphi)$ on the Bloch sphere. The spinor lives on the gray circle parametrized by the angles γ and φ.

and find

$$\cot\delta = \frac{\beta B - \lambda\cos\alpha}{\lambda\sin\alpha} \equiv x, \quad \text{with } \mathrm{sgn}(\delta) = \mathrm{sgn}(\lambda\sin\alpha)s. \tag{14.17}$$

Inserting eq. (14.16) into eq. (14.13) and further into (14.11) we obtain

$$\psi(\varphi) = \frac{1}{\sqrt{2\pi}} e^{i\ell\varphi} \begin{pmatrix} \chi_1 \\ e^{i\varphi}\chi_2 \end{pmatrix} = \frac{1}{\sqrt{2\pi}} e^{i\ell\varphi} \begin{pmatrix} \cos\frac{\alpha+\delta}{2} \\ e^{i\varphi}\sin\frac{\alpha+\delta}{2} \end{pmatrix}. \tag{14.18}$$

The angle δ describes the deviation of the spin direction from the direction of the magnetic field. The total angle between the spin orientation and the z-axis is given by $\gamma = \alpha + \delta$. The angle δ as a function of $|\lambda|/\beta B$ is depicted in Fig. 14.14(a). We can see that, in general, $\delta \neq 0, \pi$ and therefore the eigenspinors are not parallel or antiparallel to the external magnetic field. Spin orientations for the various states are indicated as arrows along with the dispersion relations in Fig. 14.13. Figure 14.14(b) shows a Bloch sphere representation of $\psi(\varphi)$. The spinor is characterized by the points on the gray circle. The angle γ describes the polar angle of the spinor to the z-axis. As the particle propagates around the ring, φ changes and the spin rotates. For $\alpha = 0$, $\gamma = 0, \pi$, i.e., the state is aligned parallel or antiparallel to the z-axis. In this case, no spin rotation occurs. In the limit $|\lambda|/\beta B \to \infty$, the angle δ becomes $-\alpha$, or $\pi - \alpha$, depending on s, and $\mathrm{sgn}(\lambda)$, and the spin states are aligned with the z-axis.

Adiabatic limit. Only if $|\lambda|/\beta B \to 0$, will $\delta = 0, \pm\pi$ and the eigenspinors rotate around the ring parallel or antiparallel to the external magnetic field. This case is called the *adiabatic limit*.

In the adiabatic limit for $s = +1$, $\delta \to 0$, and the spinor

$$\psi_+(\varphi) = \frac{1}{\sqrt{2\pi}} e^{i\ell\varphi} \begin{pmatrix} \cos\frac{\alpha}{2} \\ e^{i\varphi}\sin\frac{\alpha}{2} \end{pmatrix} \tag{14.19}$$

is aligned parallel to the magnetic field. In the adiabatic limit for $s = -1$,

$\delta \to \pm\pi$, and the spinor

$$\psi_-(\varphi) = \frac{1}{\sqrt{2\pi}} e^{i\ell\varphi} \begin{pmatrix} \cos\frac{\pi-\alpha}{2} \\ e^{i(\varphi+\pi)} \sin\frac{\pi-\alpha}{2} \end{pmatrix}$$

is aligned antiparallel to the magnetic field for all angles φ.

Interference in the adiabatic limit. In order to see Berry's phase appearing in the interference of partial waves, we approximate the dispersion relation (14.15) in the adiabatic limit by expanding to first order in the small parameter $|\lambda|/\beta B$ and obtain

$$\epsilon^{(\mathrm{ad})}_{\lambda,s} = \epsilon_{\lambda,s} = \left(\lambda - \frac{1}{2} s \cos\alpha\right)^2 + \frac{1}{4}\sin^2\alpha + s\beta B.$$

The dispersion is in the adiabatic limit approximated by two parabolae vertically offset in energy by the Zeeman splitting $2\beta B$, and horizontally by $\cos\alpha$, as shown in Fig. 14.15. The latter term will turn into Berry's geometrical phase in the interference, as we will see below.

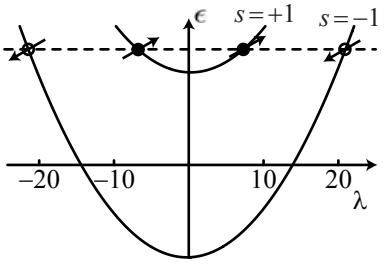


Fig. 14.15 Dispersion $\epsilon^{(\mathrm{ad})}_{\lambda,s}$ as a function of λ for $\alpha = \pi/3$ and $\beta B = 200$. At a particular energy ϵ, an electron can propagate in four distinct states. The spin orientation of these states at $\varphi = 0$ is indicated by arrows.

In order to find the transmission probability through the ring we now argue in analogy to the previous discussion of the Aharonov–Bohm effect. If the system is an isolated ring, the eigenspinors have to be periodic in φ and therefore ℓ is bound to be an integer number.

If the ring is open to leads attached at $\varphi = 0$ and $\varphi = \pi$, ℓ is not required to be integer. An electron at energy ϵ can have the four angular momentum values

$$\ell^{(\tau)}_s = \tau k_s r_0 - \frac{\phi}{\phi_0} - \frac{1}{2}(1 - s\cos\alpha), \tag{14.20}$$

where $\tau = \pm 1$, $s = \pm 1$, and

$$k_s r_0 = \sqrt{\epsilon - \frac{1}{4}\sin^2\alpha - s\beta B}.$$

These four states are indicated in Fig. 14.15. Note that the group velocity of a particular state is given by $v_{\mathrm{G}} = \tau\hbar k_s/m$.

States of a particular spin entering the ring from a lead at $\varphi = 0$ at a particular energy ϵ are split into two partial waves propagating in different directions. For example, an electron entering the ring at $\varphi = 0$ in the spin-state (14.19), i.e., $(\cos(\alpha/2), \sin(\alpha/2))$ (we call it $s = +1 \equiv \uparrow$ for convenience) can traverse the ring in a clockwise direction ($\tau = 1$), or in a counterclockwise direction ($\tau = -1$). The corresponding pair of ℓ-states is indicated in Fig. 14.15 by filled circles. The transmission amplitude for ending up in the $\uparrow$-state is in the first case

$$t^{(+1)}_{\uparrow\uparrow} = \frac{1}{2} e^{i\ell^{(+1)}_{+1}\pi} \begin{pmatrix} \cos\left(\frac{\alpha}{2}\right) \\ \sin\left(\frac{\alpha}{2}\right) e^{i\pi} \end{pmatrix} \begin{pmatrix} \cos\left(\frac{\alpha}{2}\right) \\ \sin\left(\frac{\alpha}{2}\right) \end{pmatrix} = \frac{1}{2} e^{i\ell^{(+1)}_{+1}\pi} \cos\alpha,$$

and in the second case

$$t^{(-1)}_{\uparrow\uparrow} = \frac{1}{2} e^{-i\ell^{(-1)}_{+1}\pi} \begin{pmatrix} \cos\left(\frac{\alpha}{2}\right) \\ \sin\left(\frac{\alpha}{2}\right) e^{-i\pi} \end{pmatrix} \begin{pmatrix} \cos\left(\frac{\alpha}{2}\right) \\ \sin\left(\frac{\alpha}{2}\right) \end{pmatrix} = \frac{1}{2} e^{-i\ell^{(-1)}_{+1}\pi} \cos\alpha.$$

The total transmission probability for an electron entering in the ↑-state to leave in the same spin state is therefore

$$\mathcal{T}_{\uparrow\uparrow} = \frac{1}{4}\cos^2\alpha \left| e^{i\ell_{+1}^{(+1)}\pi} + e^{-i\ell_{+1}^{(-1)}\pi} \right|^2 = \cos^2\alpha \cos^2 \frac{(\ell_{+1}^{(+1)} + \ell_{+1}^{(-1)})\pi}{2}.$$

The factor $(\ell_{+1}^{(+1)} + \ell_{+1}^{(-1)})/2$ appearing in the phase of the second cosine is the arithmetic mean of the two interfering ℓ-values. Because they lie at the same energy but at opposite wings of the same parabola (cf. Fig. 14.15), the mean value is the ℓ-value, where the parabola has its minimum, i.e., at $\ell = -\phi/\phi_0 - (1 - \cos\alpha)/2$.

In a similar way we find the transmission probability for an electron entering in the ↑-state to leave in the orthogonal ↓-state $(\sin(\alpha/2), -\cos(\alpha/2))$

$$\mathcal{T}_{\downarrow\uparrow} = \sin^2\alpha \cos^2 \frac{(\ell_{+1}^{(+1)} + \ell_{+1}^{(-1)})\pi}{2},$$

for an electron in the ↓-state to leave in the ↑-state

$$\mathcal{T}_{\uparrow\downarrow} = \cos^2\alpha \cos^2 \frac{(\ell_{-1}^{(+1)} + \ell_{-1}^{(-1)})\pi}{2},$$

and for an electron in the ↓-state to leave in the same state

$$\mathcal{T}_{\downarrow\downarrow} = \sin^2\alpha \cos^2 \frac{(\ell_{-1}^{(+1)} + \ell_{-1}^{(-1)})\pi}{2}.$$

In the latter two expressions, the term $(\ell_{-1}^{(+1)} + \ell_{-1}^{(-1)})/2$ can again be read in principle from Fig. 14.15 to be the position of the minimum at $\ell = -\phi/\phi_0 - (1 + \cos\alpha)$. The total transmission is then given by the sum of these four transmission channels, i.e., by

$$\begin{aligned}
\mathcal{T} &= \mathcal{T}_{\uparrow\uparrow} + \mathcal{T}_{\downarrow\downarrow} + \mathcal{T}_{\downarrow\uparrow} + \mathcal{T}_{\uparrow\downarrow} \\
&= \cos^2 \frac{(\ell_{+1}^{(+1)} + \ell_{+1}^{(-1)})\pi}{2} + \cos^2 \frac{(\ell_{-1}^{(+1)} + \ell_{-1}^{(-1)})\pi}{2} \qquad (14.21) \\
&= 1 + \frac{1}{2}\left\{ \cos\left[2\pi\frac{\phi}{\phi_0} + \pi(1-\cos\alpha) \right] + \cos\left[2\pi\frac{\phi}{\phi_0} - \pi(1-\cos\alpha) \right] \right\} \\
&= 1 + \cos\left[\pi(1-\cos\alpha)\right] \cos\left[2\pi\frac{\phi}{\phi_0} \right]. \qquad (14.22)
\end{aligned}$$

The transmission is modulated by the phase factor

$$\Delta\varphi_{\mathrm{B}} = \pi(1 - \cos\alpha) \qquad (14.23)$$

in addition to the Aharonov–Bohm phase. It is called Berry's phase. The value of Berry's phase has the following interpretation. As the spin moves around the ring it experiences a magnetic field that rotates around the z-axis. The rotating magnetic field vector spans a solid angle $2\pi(1 - \cos\alpha)$. In our example, Berry's phase is exactly half this solid angle, i.e., it has a very simple geometric meaning. For this reason,

Berry's phase is often called a *geometric phase*. It contains geometric information about parameter space history.

An experimental consequence of Berry's phase in an Aharonov–Bohm type of experiment would be an amplitude modulation in the oscillatory magnetoconductance. This can be seen by realizing that in an experiment where B_z is varied, while B_r remains constant, the angle α, and thereby Berry's phase, changes. A node of the Aharonov–Bohm oscillations would be expected at $\alpha = \pi/3$, or $B_r = \sqrt{3}B_z$, because there, in eq. (14.22) the prefactor containing Berry's phase vanishes. This is illustrated in Fig. 14.16 assuming a radius $r_0 = 500\,\mathrm{nm}$ and an in-plane radial field $B_r = 100\,\mathrm{mT}$.

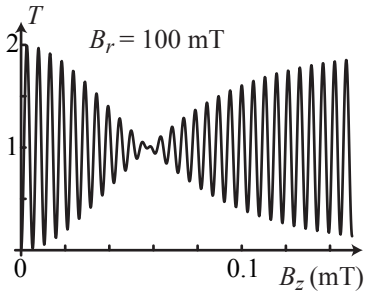


Fig. 14.16 Amplitude modulation of Aharonov–Bohm oscillations arising according to eq. (14.22) as a result of an in-plane radial magnetic field $B_r = 100\,\mathrm{mT}$ in a ring with radius 500 nm.

Spin in a rotating magnetic field. The above-mentioned geometric meaning appears in a related problem, where we consider the time evolution of a spin in a rotating magnetic field

$$\mathbf{B}(t) = B(\sin\alpha\cos\omega_0 t, \sin\alpha\sin\omega_0 t, \cos\alpha).$$

This situation is schematically depicted in Fig. 14.17. The spin evolution is governed by the hamiltonian

$$H(t) = \frac{1}{2}g\mu_\mathrm{B}\mathbf{B}(t)\boldsymbol{\sigma} = \frac{1}{2}g\mu_\mathrm{B}B\begin{pmatrix}\cos\alpha & \sin\alpha e^{-i\omega_0 t}\\ \sin\alpha e^{i\omega_0 t} & -\cos\alpha\end{pmatrix}.$$

Before we tackle the time-dependent problem, we regard the time t as a parameter and determine the eigenstates and eigenvalues of $H(t)$. Using the *Ansatz* [note the similarity to eq. (14.11)]

$$\psi(t) = \begin{pmatrix}\chi_1\\ \chi_2 e^{i\omega_0 t}\end{pmatrix}$$

we obtain the equation

$$\frac{1}{2}g\mu_\mathrm{B}B\begin{pmatrix}\cos\alpha & \sin\alpha\\ \sin\alpha & -\cos\alpha\end{pmatrix}\begin{pmatrix}\chi_1\\ \chi_2\end{pmatrix} = E\begin{pmatrix}\chi_1\\ \chi_2\end{pmatrix}$$

for a spin in a static magnetic field at tilt angle α to the z-axis. As a result we obtain the two eigenvectors of $H(t)$ given by

$$\psi_+(t) = \begin{pmatrix}\cos\frac{\alpha}{2}\\ \sin\frac{\alpha}{2}e^{i\omega_0 t}\end{pmatrix},\quad \psi_-(t) = \begin{pmatrix}-\sin\frac{\alpha}{2}\\ \cos\frac{\alpha}{2}e^{i\omega_0 t}\end{pmatrix}$$

with time-independent energies $E_\pm = \pm g\mu_\mathrm{B}B/2$.

We now return to the time-dependent problem. If the external magnetic field rotates slow enough, the spin can adiabatically follow the field. We therefore expand the solution of

$$i\hbar\partial_t\psi(t) = H(t)\psi(t)$$

in the adiabatic eigenstates, i.e.,

$$\psi(t) = c_1(t)\psi_+(t) + c_2(t)\psi_-(t),$$

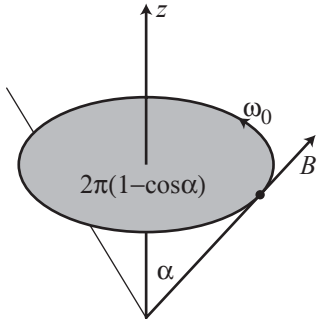


Fig. 14.17 The magnetic field vector revolves around the z-axis at constant angular velocity ω_0 while the angle α remains constant. By its motion it spans a solid angle $2\pi(1-\cos\alpha)$.

and obtain the equation for the coefficients

$$i\hbar\partial_t \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \frac{1}{2}\hbar\omega_0 \begin{pmatrix} 1+\beta B-\cos\alpha & \sin\alpha \\ \sin\alpha & 1-\beta B+\cos\alpha \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

where $\beta = g\mu_{\mathrm{B}}/\hbar\omega_0$. Note that the effective hamiltonian matrix on the right-hand side is now time-independent. We now let $(c_1, c_2) = (A_1, A_2)\exp(-iEt/\hbar)$ and obtain

$$\begin{pmatrix} 1+\beta B-\cos\alpha-\epsilon & \sin\alpha \\ \sin\alpha & 1-\beta B+\cos\alpha-\epsilon \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = 0,$$

with $\epsilon = 2E/\hbar\omega_0$. There is a striking similarity to the eigenvalue problem in eq. (14.14). Indeed, the problems become identical for $\lambda = 1$, if we absorb the energy offset $1/4$ in ϵ. We can therefore directly read the eigenenergies ϵ, and the eigenspinors (A_1, A_2) from the previous problem of the ring in a textured magnetic field. We can also state that the adiabatic limit is reached for $\beta B \to \infty$, i.e., when the energy $\hbar\omega_0$ is very small compared to the Zeeman splitting $g\mu_{\mathrm{B}}B$. This is equivalent to saying that a superposition of up- and downspin (with respect to the magnetic field direction at a particular time) oscillates a very large number of times within the period of the rotation of the magnetic field. In the adiabatic limit, the energy eigenvalues are given by

$$\epsilon_s^{(\mathrm{ad})} = s\beta B + 1 - s\cos\alpha.$$

Assume that the spin starts at $t = 0$ in a state where it is aligned with the magnetic field. The adiabatic time evolution of the state is then given by

$$\psi(t) = e^{-i[g\mu_{\mathrm{B}}Bt/2\hbar+\omega_0 t(1-\cos\alpha)/2]} \begin{pmatrix} \cos\frac{\alpha}{2} \\ \sin\frac{\alpha}{2}e^{i\omega_0 t} \end{pmatrix}.$$

At time $t = 0$ the state is

$$\psi(0) = \begin{pmatrix} \cos\frac{\alpha}{2} \\ \sin\frac{\alpha}{2} \end{pmatrix}.$$

After a time period $T = 2\pi/\omega_0$ of one full revolution of the magnetic field vector, the state has changed into

$$\psi(t) = e^{-i[\pi g\mu_{\mathrm{B}}B/\hbar\omega_0+\pi(1-\cos\alpha)]} \begin{pmatrix} \cos\frac{\alpha}{2} \\ \sin\frac{\alpha}{2} \end{pmatrix},$$

i.e., it has acquired a *dynamic* phase $\Delta\varphi_{\mathrm{D}} = \pi g\mu_{\mathrm{B}}B/\hbar\omega_0$ given by the state's energy and the elapsed time and the *geometric* phase $\Delta\varphi_{\mathrm{B}}$ from eq. (14.23) given by half the solid angle spanned by the rotating magnetic field vector. A similar argument for a state starting antiparallel to the magnetic field leads to the same dynamic and geometric phases multiplied with -1.

Experiments. The experimental observation of Berry's phase in semiconducting ring structures has proven to be challenging, and has remained an open problem until today. Attempts have been made to realize the required magnetic field texture by placing a small magnetic disk in the center of ring structures (Jacobs and Giordano, 1998; Ye *et al.*, 1999). Other attempts exploit the spin–orbit interaction relevant in semiconductors, such as InAs, or p-GaAs. We will discuss these attempts in the following section.

14.5 Aharonov–Casher phase and spin–orbit interaction induced phase effects

Aharonov and Casher have pointed out that an electromagnetic dual to the Aharonov–Bohm effect exists (Aharonov and Casher, 1984). It is realized when a *neutral* particle with magnetic moment (spin) encircles a line of charge (charge density λ) creating a radial electric field, as depicted in Fig. 14.3(b). In this case, circular motion of the magnetic moment μ also leads to the accumulation of a quantum phase, the so-called Aharonov–Casher phase (the existence of this phase effect was, however, pointed out earlier, in Anandan, 1982). Early attempts to measure this effect were performed with neutrons (Cimmino *et al.*, 1989).

The effect can be understood by considering the relativistic transformation of the electric field into the moving reference frame of the neutral particle. The magnetic moment experiences a magnetic field $\mathbf{B} = c^{-2}\mathbf{v}\times\mathbf{E}$ brought about by this transformation, and changes its orientation under the influence of this field. The analogy to the Aharonov–Bohm phase becomes obvious when we write down the canonical momentum of the particle in the electric field $\mathbf{E}$. It is given by

$$\mathbf{p} = m\mathbf{v} + \frac{1}{c}\boldsymbol{\mu}\times\mathbf{E}$$

showing a similar form to the canonical momentum $\mathbf{p} = m\mathbf{v} + q\mathbf{A}$ of a particle with charge q in the presence of a magnetic vector potential $\mathbf{A}$. The line charge creates an electric field $\mathbf{E} = \mathbf{e}_r\lambda/2\pi\epsilon_0 r_0$ at distance r_0, where $\mathbf{e}_r$ is a unit vector pointing in a radial direction. The Aharonov–Casher phase is then given by

$$\Delta\varphi_{\mathrm{AC}} = \frac{1}{c}\oint \boldsymbol{\mu}\times\mathbf{E}d\mathbf{s} = \frac{\mu_z\lambda}{c\epsilon_0}.$$

As in the case of the Aharonov–Bohm phase, the Aharonov–Casher phase is independent of the radius r_0 of the orbit.

Although the original setting envisioned by Aharonov and Casher cannot be easily realized in semiconductor nanostructures because electrons and holes are *charged* particles, the notion of the Aharonov–Casher phase has in the literature of recent years been sometimes extended to the case of *charged* particles with spin moving in *arbitrarily* oriented electric fields

in solids. Compared to the case of a neutral particle, the presence of the charge leads to a correction of the magnetic field experienced by the particle by a factor of 1/2, the so-called Thomas factor. Allowing for arbitrary electric field orientation makes the situation equivalent to the presence of spin–orbit interaction in a solid. Using this rather extended notion of the Aharonov–Casher phase, any influence of spin–orbit interaction on electron (or hole) interference can be seen as a manifestation of this phase. Rather than stretching the meaning of the Aharonov–Casher phase far beyond the original context, we prefer to call such effects *spin–orbit interaction induced* in the following.

In order to get to the roots of spin–orbit interaction induced phase effects, we consider the circular motion of an electron in the presence of Rashba spin–orbit interaction (see section 9.6). The ultimate goal is to find the transmission (in lowest order) through an open two-terminal ring including the interference effects arising from clockwise and counterclockwise partial waves, in full analogy with the previous discussion of the Aharonov–Bohm effect. We will discuss the problem in two steps. In the first step we discuss the case where no external magnetic field (or flux) is applied. In the second step such an additional magnetic field is included and an interplay between Aharonov–Bohm effect, Zeeman splitting and spin–orbit interaction gives rise to qualitatively new behavior of the system.

Step I: Circular motion without magnetic field. The two-dimensional electronic motion in the presence of Rashba spin–orbit interaction can be described by the hamiltonian [cf., eq. (9.6)]

$$H = \frac{\mathbf{p}^2}{2m} + \frac{\alpha}{\hbar}\boldsymbol{\sigma}(\mathbf{p}\times\mathbf{E}) + V(r),$$

where $\mathbf{E}$ is oriented in the positive z-direction and describes the average electric field experienced by the electrons, $V(r)$ is the radial confinement potential, and the motion in the z-direction has been assumed to be separable. In cylindrical coordinates this hamiltonian reads

$$H = -\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}\right] + V(r) - \frac{\hbar^2}{2mr^2}\frac{\partial^2}{\partial\varphi^2} - i\alpha_{\mathrm{R}}\left[(\sigma_x\sin\varphi - \sigma_y\cos\varphi)\frac{\partial}{\partial r} + (\sigma_x\cos\varphi + \sigma_y\sin\varphi)\frac{1}{r}\frac{\partial}{\partial\varphi}\right],$$

with $\alpha_{\mathrm{R}} = \alpha\langle E_z\rangle$. Care has to be taken to find the correct form of the hamiltonian for a strictly one-dimensional ring. The conventional procedure of neglecting the radial derivatives and letting $r = r_0$ fails here (Meijer *et al.*, 2002). The correct form of the one-dimensional hamiltonian is

$$H = -\frac{\hbar^2}{2mr_0^2}\frac{\partial^2}{\partial\varphi^2} - \frac{i\alpha_{\mathrm{R}}}{r_0}\left[-\frac{1}{2}(\sigma_x\sin\varphi - \sigma_y\cos\varphi) + (\sigma_x\cos\varphi + \sigma_y\sin\varphi)\frac{\partial}{\partial\varphi}\right]. \quad (14.24)$$

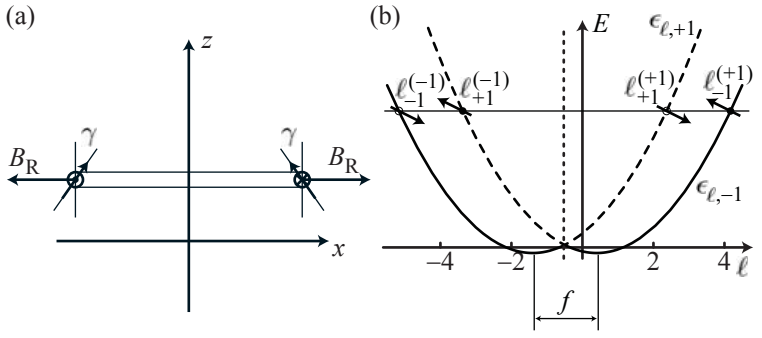


Fig. 14.18 (a) Cross-sectional view of a ring with an electron propagating in the counterclockwise direction. The effective Rashba field and the spin orientation are indicated. (b) Dispersion relation for an electron on a one-dimensional ring in the presence of Rashba spin–orbit interaction. The dotted line is defined by $\ell = -1/2$. Open and closed circles indicate pairs of interfering angular momentum states propagating all at the same energy. The arrows indicate the spin orientation for $\varphi = 0$ in the x-z plane of the Bloch sphere.

Here we introduce the parameter $Q_{\rm R} = 2m\alpha_{\rm R}r_0/\hbar^2$ which characterizes the relative strength of the spin–orbit interaction effects. They are absent for $Q_{\rm R} = 0$ and very strong for $Q_{\rm R} \gg 1$.

Inserting the spinor (14.11) into the one-dimensional eigenvalue problem gives the following equation for the amplitudes $\chi_{1,2}$ (Frustaglia and Richter, 2004):

$$\begin{pmatrix} (\lambda-\frac{1}{2})^2-\epsilon & Q_{\rm R}\lambda \\ Q_{\rm R}\lambda & (\lambda+\frac{1}{2})^2-\epsilon \end{pmatrix} \begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} = 0, \tag{14.25}$$

where the eigenenergy is $E = \hbar^2\epsilon/2mr_0^2$, and $\lambda = \ell+1/2$. This equation is identical to eq. (14.12) for the ring in a textured magnetic field, if we let $B_z = 0$, and if we identify

$$B_r = Q_{\rm R}\lambda/\beta \equiv B_{\rm R}$$

to be the effective radial magnetic field, also called the Rashba field $B_{\rm R}$. This makes it clear that the problem of the ring with Rashba interaction is very similar to the problem of the ring in a textured magnetic field. However, an important difference is that the radial field component is created dynamically by the motion of the electron. It is therefore proportional to the angular momentum of the electron via $\lambda = \ell + 1/2$. For $\ell > -1/2$ the field points radially outwards, while for $\ell < -1/2$ the field points radially inwards. However, since we have no field component along z, the angle α between the total magnetic field vector and the z-axis is $\mathrm{sgn}(\lambda)\pi/2$, independent of the magnitude of the angular momentum. The eigenspinors are generally not aligned with the direction of the Rashba field. This situation is schematically depicted in Fig. 14.18(a) which shows a cross-sectional view through the ring with the Rashba field and the spin orientation for an electron with positive angular momentum. In the limit $Q_{\rm R} \to \infty$, implying either r_0 going to infinity (motion along a straight line), or the Rashba coefficient $\alpha_{\rm R}$ going to infinity (very strong spin–orbit interaction), the spin is aligned with the Rashba field. This is the adiabatic limit, because the spin follows the field adiabatically when the state propagates around the ring.

Inserting $B_z = 0$ and $B_r = B_\mathrm{R}$ in eq. (14.15) gives the exact eigenenergies

$$\epsilon_{\ell,s} = \left[\ell + \frac{\Phi_\mathrm{SO}^{(\tau,s)}}{2\pi}\right]^2 + \frac{1}{4}(1-f^2), \tag{14.26}$$

where the spin–orbit induced phase $\Phi_\mathrm{SO}^{(\tau,s)}$ is given by

$$\Phi_\mathrm{SO}^{(\tau,s)} = \pi\,[1 + s\tau f]$$

with $\tau = \mathrm{sgn}(\ell + 1/2) = \pm 1$, $f = \sqrt{1+Q_\mathrm{R}^2}$, and $s = \pm 1$ identifying the two spin eigenstates. The two spin branches of the dispersion ($s = \pm 1$) are schematically shown in Fig. 14.18(b). They touch at $\ell = -1/2$ at the energy $E = \hbar^2/8mr_0^2$. Equation (14.26) is similar to the dispersion relation of the Aharonov–Bohm ring (14.5) with the Aharonov–Bohm flux ϕ/ϕ_0 replaced by $\Phi_\mathrm{SO}^{(\tau,s)}$. The dispersion is essentially composed of two parabolae shifted in ℓ by different amounts. The two minima of the dispersion parabolae occur at $\ell = -1/2 \pm \sqrt{1+Q_\mathrm{R}^2}/2$. For zero Rashba spin–orbit interaction, one minimum is at $\ell = 0$, and the other is at $\ell = -1$. For very large Q_R, the minima occur at approximately $\pm Q_\mathrm{R}$. The sign of the group velocity given by eq. (14.6) gives the direction of propagation of the states. States ℓ for which $\partial E/\partial\ell > 0$ (< 0) propagate counterclockwise (clockwise) around the ring. At any particular energy E, four states form two pairs travelling in clockwise or counterclockwise direction, respectively.

The eigenspinors are given by eq. (14.18) with $\cot\delta = \mathrm{sgn}(\lambda)Q_\mathrm{R}$ and $\mathrm{sgn}(\delta) = \mathrm{sgn}(\lambda)s$. As a result, the angle $\gamma = \pi/2 + \delta$ between the spin and the z-axis does not depend on the magnitude of λ (or ℓ), but only on its sign. An incoming beam spin-polarized at the angle γ can always be split in two counterpropagating partial waves. A derivation of the transmission along the lines that were followed in section 14.4 leads in analogy to eq. (14.21) to

$$\mathcal{T} = \cos^2\frac{(\ell_{+1}^{(+1)} + \ell_{-1}^{(-1)})\pi}{2} + \cos^2\frac{(\ell_{-1}^{(+1)} + \ell_{+1}^{(-1)})\pi}{2}$$

with

$$\ell_s^{(\tau)} = \tau k r_0 - \frac{\Phi_\mathrm{SO}^{(\tau,s)}}{2\pi}.$$

Inserting these angular momentum values into the expression for the transmission gives the final result

$$\mathcal{T} = 1 - \cos\left(\pi\sqrt{1+Q_\mathrm{R}^2}\right).$$

The transmission is modulated by the strength Q_R of the spin–orbit interaction as shown in Fig. 14.19. Tuning the Rashba interaction strength α_R changes Q_R. As a consequence, the interference can be tuned from destructive to constructive and vice versa.

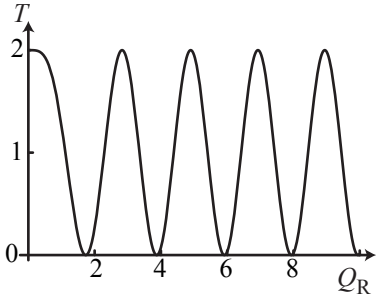


Fig. 14.19 Modulation of the transmission probability T as a function of the Rashba spin–orbit interaction parameter Q_R.

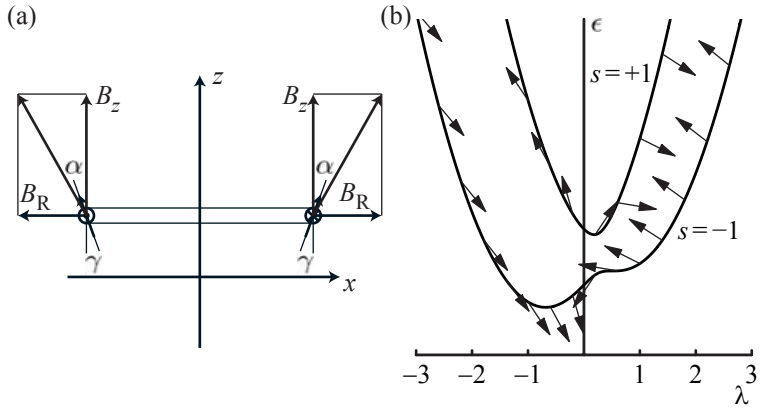


Fig. 14.20 (a) Cross-sectional view of a ring with an electron propagating in the counterclockwise direction. The effective Rashba field $B_{\rm R}$ and the externally applied field B_z add to the total effective magnetic field which is at an angle α to the z-axis. The spin orientation differs from the direction of the total magnetic field and encloses an angle γ with the z-axis. (b) Dispersion relation for an electron on a one-dimensional ring in the presence of Rashba spin–orbit interaction and a perpendicular magnetic field. The arrows indicate the direction of the spin at $\varphi = 0$. The dispersion was calculated for $Q_{\rm R} = 1$ and $\beta B_z = 0.5$.

Step II: Circular motion with magnetic field. In experiments on ring structures, magnetic fields normal to the plane of the ring are usually applied in order to see Aharonov–Bohm type of oscillations. The field has two distinct effects: it introduces an Aharonov–Bohm phase in the orbital motion of the electron, and it leads to a Zeeman effect of the electron spin. The previous model of the one-dimensional ring can be extended to incorporate both effects (Frustaglia and Richter, 2004). The hamiltonian (14.24) is extended to

$$H = \frac{\hbar^2}{2mr_0^2}\left(-i\frac{\partial}{\partial\varphi} + \frac{\phi}{\phi_0}\right)^2 + \frac{1}{2}g\mu_{\rm B}B\sigma_z$$
$$+\frac{\alpha_{\rm R}}{r_0}\left[\frac{i}{2}(\sigma_x\sin\varphi - \sigma_y\cos\varphi) + (\sigma_x\cos\varphi + \sigma_y\sin\varphi)\left(-i\frac{\partial}{\partial\varphi} + \frac{\phi}{\phi_0}\right)\right].$$

Inserting the one-dimensional eigenspinors (14.11) leads to the matrix equation

$$\begin{pmatrix} (\lambda - \frac{1}{2})^2 + \beta B_z - \epsilon & Q_{\rm R}\lambda \\ Q_{\rm R}\lambda & (\lambda + \frac{1}{2})^2 - \beta B_z - \epsilon \end{pmatrix}\begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} = 0, \quad (14.27)$$

where $\epsilon_{\rm Z} = mr_0^2 g\mu_{\rm B}B/\hbar^2$, $\lambda = \ell + \phi/\phi_0 + 1/2$, and $\beta = g\mu_{\rm B}mr_0^2/\hbar^2$. Comparing with eq. (14.12) for the ring with Zeeman interaction in a textured magnetic field shows that the spin is here moving in a superposition of the angular momentum-dependent Rashba field

$$B_{\rm R} = \frac{Q_{\rm R}}{\beta}(\ell + \frac{\phi}{\phi_0} + \frac{1}{2}),$$

and the external field B_z normal to the plane of the ring. This situation is schematically depicted in Fig. 14.20(a). The total magnetic field is at the angle α to the z-axis with

$$\tan\alpha = \frac{B_{\rm R}}{B_z} = \frac{Q_{\rm R}(\ell + \frac{\phi}{\phi_0} + \frac{1}{2})}{\beta B_z}.$$

The eigenvalue solutions of the above problem are [cf., eq (14.15)]

$$\epsilon_{\lambda,s} = \lambda^2 + \frac{1}{4} + s\sqrt{Q_{\rm R}^2\lambda^2 + (\lambda - \beta B_z)^2}, \tag{14.28}$$

where $s = \pm 1$ denotes the two spin states. An example of this dispersion relation is depicted in Fig. 14.20(b). The Aharonov–Bohm effect enters the dispersion merely as a shift of the dispersion along ℓ, not visible in the plot vs. λ. The Zeeman effect leads to the Zeeman gap $2\beta B_z$ at $\lambda = 0$ removing the degeneracy arising at $B_z = 0$. As in the previous discussion without external magnetic field, the spin is given by eq. (14.18). It is usually not aligned with the total magnetic field. The spin orientation at $\varphi = 0$ is indicated for the states in the dispersion shown in Fig. 14.20(b).

For the angle δ describing the deviation of the spin direction from the total effective magnetic field we obtain from eq. (14.17)

$$\cot\delta = \frac{1}{Q_{\rm R}}\left(\frac{\beta B_z}{\lambda} - \frac{1}{2}\right)^2 + Q_{\rm R} - \frac{1}{4Q_{\rm R}} \quad \text{with } \mathrm{sgn}(\delta) = s.$$

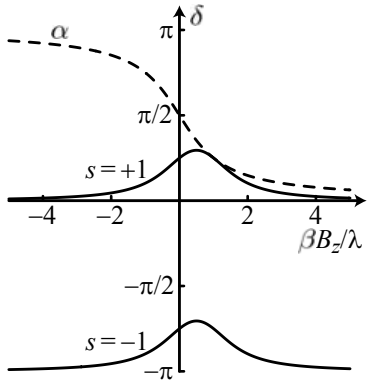


Fig. 14.21 The angle δ as a function of $\beta B_z/\lambda$ calculated for $Q_{\rm R} = 1$ and $s = \pm 1$. The angle α is plotted as a reference.

The angle δ depends on the parameters $\beta B_z/\lambda$ and $Q_{\rm R}$. Figure 14.21 shows a plot as a function of $\beta B_z/\lambda$. The angle α is included for comparison. The angle δ, i.e., the deviation of the spin orientation from the effective magnetic field direction α, has a maximum for $\beta B_z/\lambda = 1/2$. For large positive or negative $\beta B_z/\lambda$, δ tends to zero, meaning that the adiabatic limit is reached. If the angle α is close to 0 or $\pm\pi/2$, where δ is close to zero, then the Rashba interaction is of negligible influence in the adiabatic limit. If, however, δ is close to zero, where α differs appreciably from zero or $\pm\pi$, then an interesting adiabatic regime is reached, similar to the case of the ring in a magnetic field texture. In this case, Berry phase effects would be expected in an Aharonov–Bohm interference experiment.

The interference effects appearing in the transmission $\mathcal{T}$ through the open ring cannot be discussed as easily as in the previous examples. The reason is that the energy dispersion in eq. (14.28) cannot be solved analytically for λ. However, we can see from Fig. 14.20(b) that an electron at a given energy will again have the possibility of propagating along four distinct channels (two clockwise and two counterclockwise) around the ring. In contrast to the case of $B_z = 0$ discussed before, the spin orientation of a particular ℓ-state will now depend on the magnitude of ℓ, and there are no pairs of mutually aligned spins for clockwise and counterclockwise propagation at the same energy. The state of an incoming electron with a certain spin orientation will therefore propagate around the ring in a superposition of the available clockwise (or counterclockwise) states and a number of interference terms arises leading to beating-like patterns in Aharonov–Bohm oscillations. A more detailed discussion can be found in Molnar *et al.*, 2004, Frustaglia and Richter, 2004, and Aeberhard *et al.*, 2005.

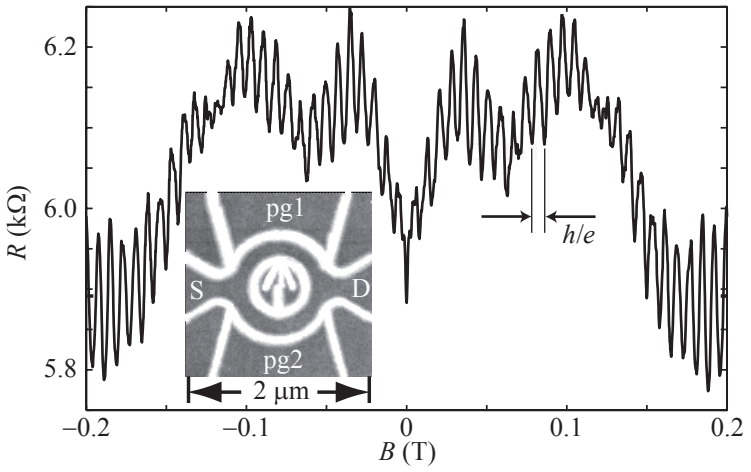


Fig. 14.22 Resistance of a quantum ring structure realized with AFM lithography on the basis of a two-dimensional hole gas in GaAs. The resistance is measured between source (S) and drain (D) as a function of the magnetic field B applied normal to the plane of the ring. (Reprinted with permission from Grbic *et al.*, 2007. Copyright 2007 by the American Physical Society.)

14.6 Experiments on spin–orbit interaction induced phase effects in rings

The unambiguous identification of spin–orbit interaction induced phase effects in ring structures is still an open and challenging experimental task. Initial experimental work has been done on n-type AlSb/InAs/AlSb quantum well structures with single two-terminal rings between 1 and 2 µm diameter and a mean free path of about 1 µm (Morpurgo *et al.*, 1998), where the Rashba term is believed to be dominant. Further work was performed on Rashba-dominated n-InGaAs-based ring structures of similar size and mean free path (Nitta *et al.*, 1999; Nitta *et al.*, 2000; Nitta *et al.*, 2002), and on arrays of rings (Bergsten *et al.*, 2006). In all these experiments, a magnetic field was applied normal to the plane of the ring leading to periodic Aharonov–Bohm type of oscillations of the conductance as a function of magnetic flux enclosed by the ring. Averaged Fourier spectra of these oscillations (Meijer *et al.*, 2004) were used in most of the experiments for identifying side bands of the main h/e Aharonov–Bohm peak. These side bands were interpreted as evidence for the influence of spin–orbit induced phase effects. Ring structures fabricated from HgTe/HgCdTe quantum wells have also been studied (Konig *et al.*, 2006). Beating-like patterns in the raw Aharonov–Bohm modulated magnetoconductance of an InAs ring were observed as well (Yang *et al.*, 2004).

As an alternative to n-type systems with strong spin–orbit interaction, ring structures based on p-type GaAs material have been investigated (Yau *et al.*, 2002; Habib *et al.*, 2007; Grbic *et al.*, 2007). Generally, spin–orbit interaction effects are expected to be more pronounced in the valence than in the conduction band. Figure 14.22 shows the measure-

ment of the oscillatory magnetoresistance of a quantum ring realized with AFM lithography on the basis of a two-dimensional hole gas (Grbic *et al.*, 2007). The oscillations are superimposed on a slowly varying background. Although there is a beating-like pattern in the oscillations, it is not straightforward to identify spin–orbit related effects unambiguously in the raw data, because even conventional n-type GaAs samples, where spin–orbit effects are expected to be not important, can show similar beating effects.

14.7 Decoherence

Decoherence is a general phenomenon describing the way that the phase difference $\Delta\varphi = \varphi(\mathbf{r}, t) - \varphi(\mathbf{r}', t')$ between an electron wave at position $\mathbf{r}$ and time t and the same wave at position $\mathbf{r}'$ and time t' gets lost, if the time-difference $|t - t'|$ becomes too large. This phenomenon plays an important role if one tries to understand the transition from the quantum to the classical description of nature, because it implies the loss of interference phenomena which do not occur in classical mechanics. There are two very basic ways to understand decoherence. On the one hand, decoherence of a quantum system can be seen as the result of entanglement between the states of this system and its environment (or 'bath'). On the other hand, it can be seen as the interaction of the system with a fluctuating potential (or field) that is created by the environment. This fluctuating field leads to statistical, i.e., random, phase changes of the system's wave function and therefore the phase information important for the observation of interference is lost.

14.7.1 Decoherence by entanglement with the environment

As a model system showing decoherence, we consider a quantum ring that is very weakly coupled to a bath of quantum particles with many degrees of freedom. In our example, this bath is given by all the electrons populating states in the ring in thermodynamic equilibrium, and we consider their interaction with a single (nonequilibrium) electron injected from one of the contacts. In general, the bath could also be a system of photons or phonons, but it turns out that, in low-temperature experiments (below 4.2 K), the interactions with the photonic or phononic environment can often be neglected.

The states of the injected electron in the ring are described by the orthonormal basis functions $\varphi_n(x)$, where n is a set of quantum numbers allowing us to distinguish states in the left and the right arm of the interferometer. For the sake of simplicity, we use the labels $n = \ell$ (left arm of the ring) and $n = r$ (right arm of the ring). The coordinates x describe the position of the electron. The states of the environment, i.e., of the system of all other electrons in the ring, are described in the orthonormal basis $\chi_\alpha(\eta)$ with quantum numbers α and coordinates η.

We assume that at time zero the system is the product state

$$\psi(t=0) = [\varphi_\ell(x) + \varphi_r(x)]\,\chi_0(\eta).$$

If the total state of the electron plus environment is described by such a product wave function, we say that the two systems are not entangled. Over time the system will evolve and the wave function at time t will have the general form

$$\psi(t) = \sum_{n\alpha} c_{n\alpha}(t)\varphi_n(x)\chi_\alpha(\eta),$$

which is no longer the product of a state of the system and the environment. In this case we talk about an entangled state meaning that the system and the environment are entangled. Using this wave function we find for the probability density

$$\begin{aligned}
|\psi(t)|^2 &= \sum_{n\alpha,m\beta} c_{n\alpha}(t)c^\star_{m\beta}(t)\varphi_n(x)\varphi^\star_m(x)\chi_\alpha(\eta)\chi^\star_\beta(\eta) \\
&= \sum_{n\alpha} |c_{n\alpha}(t)|^2\,|\varphi_n(x)|^2\,|\chi_\alpha(\eta)|^2 \\
&\quad + {\sum_{n\alpha,m\beta}}' c_{n\alpha}(t)c^\star_{m\beta}(t)\varphi_n(x)\varphi^\star_m(x)\chi_\alpha(\eta)\chi^\star_\beta(\eta).
\end{aligned}$$

Here the primed sum implies summation over indices $n\alpha \neq m\beta$ only, meaning that this term contains the interference effects. The coefficients $c_{n\alpha}$ depend on the time t. If we calculate the expectation value of an operator acting only on the states of the injected electron, such as the operator of the current density in the ring, $\hat{j}(x)$, we integrate out the coordinates η of all other equilibrium carriers. The interference term turns into

$$\begin{aligned}
&{\sum_{n\alpha,m\beta}}' c_{n\alpha}c^\star_{m\beta}\varphi_n(x)\varphi^\star_m(x)\int d\eta\chi_\alpha(\eta)\chi^\star_\beta(\eta) \\
&\qquad = {\sum_{n\alpha,m\beta}}' c_{n\alpha}c^\star_{m\beta}\varphi_n(x)\varphi^\star_m(x)\delta_{\alpha\beta} \\
&\qquad\qquad = {\sum_{n,m}}' \left(\sum_\alpha c_{n\alpha}c^\star_{m\alpha}\right)\varphi_n(x)\varphi^\star_m(x).
\end{aligned}$$

The products $c_{n\alpha}c^\star_{m\beta}$ are the elements of the density matrix of the total system. The sum over the environment states α in the last line reduces the density matrix to that of the injected electron. We can write this

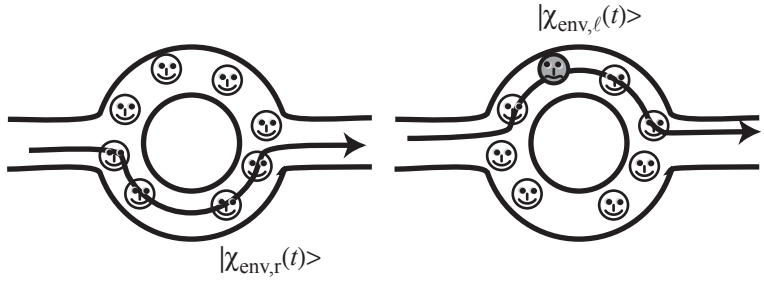


Fig. 14.23 Illustration of decoherence. The environment is represented by the smileys. Depending on the path of the electrons the resulting state of the environment differs. The environment measures the path of the electron and the interference disappears.

sum as

$$\begin{aligned}
&\sum_\alpha c_{n\alpha} c^\star_{m\alpha} \\
&= \sum_\alpha \int dx d\eta \psi^\star(x,\eta,t)\varphi_n(x)\chi_\alpha(\eta) \cdot \int dx d\eta \psi(x,\eta,t)\varphi^\star_m(x)\chi^\star_\alpha(\eta) \\
&= \sum_\alpha \int d\eta \tilde{\chi}^\star_n(\eta,t)\chi_\alpha(\eta) \cdot \int d\eta \tilde{\chi}_m(\eta,t)\chi^\star_\alpha(\eta) \\
&= \int d\eta \tilde{\chi}^\star_n(\eta,t)\tilde{\chi}_m(\eta,t).
\end{aligned}$$

Here we have performed the integration over x in the total wave function and thereby introduced effective wave functions $\tilde{\chi}_n(\eta,t)$ of the environment. Now we can write the interference term as

$$\varphi_\ell(x)\varphi^\star_r(x) \int d\eta \tilde{\chi}^\star_l(\eta,t)\tilde{\chi}_r(\eta,t) + c.c.$$

If we interpret $\chi_\ell(\eta)$ as the 'wave function' of the environment for the case that the injected electron moves through the left arm and correspondingly for $\chi_r(\eta)$, the interference term vanishes exactly if these two wave functions are orthogonal. In this case, the environment can distinguish which path the electron has taken. It is as if the environment had made a measurement on the electron and obtained information about its path. This case is schematically depicted in Fig. 14.23. The opposite extreme case arises if both wave functions of the environment are identical, such that the overlap integral is one. In this case, the environment has not acquired any information about the path of the electron and interference is not spoiled.

As a general rule we can therefore state that interference is always lost (or impaired) if information about the electron's path is acquired by the environment, i.e., if the path of the electron is 'measured' by the environment. It is important to realize that it is irrelevant whether or not this information is received by a human observer consciously performing a measurement.

14.7.2 Decoherence by motion in a fluctuating environment

We obtain a complementary view on decoherence if we consider the action of the environment on the injected electron. The environment being, for example, a bath of moving electrons, creates an interaction potential that fluctuates statistically in time and acts on the injected electron. The strength of the fluctuations will be stronger with increasing temperature because the random motion of the bath electrons will be more energetic. The fluctuating potential acts directly on the phase of the electron wave. In order to see how this phase change comes about, consider the time-dependent perturbation $V(t)$ acting on a system with the total hamiltonian

$$H = H_0 + V(t),$$

where H_0 is time-independent. Let $\psi_0(x,t)$ be a solution of the time-dependent Schrödinger equation

$$i\hbar\partial_t\psi_0 = H_0\psi_0.$$

We now try solutions of the full hamiltonian H, using a wave function of the form

$$\psi = \psi_0 e^{i\varphi(t)}$$

and find for the left-hand side of Schrödinger's equation

$$i\hbar\partial_t\psi = (i\hbar\partial_t\psi_0 - \hbar\psi_0\partial_t\varphi(t))\, e^{i\varphi(t)} = (H_0\psi_0 - \hbar\psi_0\partial_t\varphi(t))\, e^{i\varphi(t)},$$

and for the right-hand side

$$H\psi = (H_0\psi_0 + V(t)\psi_0)\, e^{i\varphi(t)}.$$

We see that

$$-\hbar\partial_t\varphi(t) = V(t), \text{ such that } \varphi(t) = -\frac{1}{\hbar}\int^t dt' V(t').$$

If a wave packet with the expectation value for the position $x(t)$ moves in a static potential $V(x)$, the additional phase is, in the semiclassical approximation, given by

$$\varphi = -\frac{1}{\hbar}\int V[x(t)]dt.$$

If the potential is not static, but created dynamically by the environment, $V(x)$ becomes an operator with a certain expectation value and an uncertainty. As a result, the phase of a partial wave also acquires an uncertainty. The influence of the environment on the partial waves consists of the factor

$$\langle e^{i\varphi}\rangle = \int P(\varphi)e^{i\varphi}d\varphi, \tag{14.29}$$

where $P(\varphi)$ is a probability distribution for the appearance of a particular phase φ. It can be shown that this phase factor is identical with the overlap of the two environment states defined above (Stern *et al.*, 1990):

$$\langle e^{i\varphi} \rangle = \int d\eta \tilde{\chi}_l^\star(\eta, t) \tilde{\chi}_r(\eta, t).$$

We can obtain a little more insight into the phase averaging described by eq. (14.29) if we assume that the total phase difference between two interfering paths is the sum of a large number of statistically independent contributions. According to the central limit theorem, the probability distribution $P(\varphi)$ is then a gaussian distribution, i.e.,

$$P(\varphi) = \frac{1}{\sqrt{2\pi \langle \delta\varphi^2 \rangle}} \exp\left(-\frac{(\varphi - \varphi_0)^2}{2\langle \delta\varphi^2 \rangle} \right).$$

The phase φ_0 is the average phase difference that is accumulated and the phase uncertainty $\langle \delta\varphi^2 \rangle$ describes the width of the distribution. Within these assumptions the integral in eq. (14.29) can be solved giving

$$\langle e^{i\varphi} \rangle = e^{i\varphi_0} e^{-\langle \delta\varphi^2 \rangle},$$

i.e., a suppression of the amplitude of the interfering phase factor $\exp(i\varphi_0)$ by the presence of the phase uncertainty. The latter is approximately given by the expression (Stern *et al.*, 1990)

$$\langle \delta\varphi^2 \rangle \approx \frac{1}{\hbar^2} \int_0^{t_0} dt \int_0^{t_0} dt' \langle V(t) V(t') \rangle, \tag{14.30}$$

where $V(t)$ represents the time-dependent potential that the electron experiences along its path between time zero (wave is split in two partial waves) and time t_0 (partial waves are brought to interference). The correlator of the potential depends only on the time difference $t - t'$, and usually it decays on the scale of a correlation time t_c that we have assumed to be much smaller than t_0 by using the central limit theorem. Assuming an exponential decay, the phase uncertainty would look like

$$\langle \delta\varphi^2 \rangle \approx \langle V^2 \rangle \frac{t_c t_0}{\hbar^2},$$

an expression which exhibits typical properties of phase uncertainty: it increases with the time t_0 which the electron spends in the interferometer, and it is proportional to the mean fluctuation amplitude $\langle V^2 \rangle$ of the interaction potential. This proportionality makes it clear that we can view the effect of decoherence as the influence of noise, or fluctuations, that exist in the environment, coupling to the system in which interference takes place.

Suppose the interfering electron propagates through an Aharonov–Bohm ring at the Fermi energy of a Fermi sea formed by all the other electrons in the ring. We can regard this electron as being coupled to the thermal bath formed by this Fermi sea via the Coulomb interaction between electrons. Local and temporal fluctuations in the electron

density will lead to a net electric potential that fluctuates in time and space, and statistically alters the phase of the interfering electron along its path. According to the Nyquist formula (fluctuation–dissipation theorem), for sufficiently high temperatures the characteristic amplitude of the fluctuation will be proportional to temperature, and therefore

$$\langle \delta\varphi^2 \rangle \propto \frac{k_\mathrm{B} T}{\hbar} t_0.$$

In the opposite limit of low temperatures the phase uncertainty and therefore the decoherence are exponentially suppressed.

In general, the details of the correlator entering eq. (14.30) depend on the specific system under investigation. Relevant input are the geometry and dimensionality of the system, the nature of the interaction, and often the temperature. Frequently, a (temperature-dependent) phase coherence time $\tau_\varphi(T)$ is introduced via the relation

$$\frac{1}{2}\langle \delta\varphi^2 \rangle = \frac{t_0}{\tau_\varphi(T)}.$$

In an ideal quantum ring, the injected electron moves ballistically with the Fermi velocity $v_\mathrm{F} = \hbar k_\mathrm{F}/m^\star$. Therefore, we can define a phase-coherence length

$$l_\varphi(T) = v_F \tau_\varphi(T) \tag{14.31}$$

and the decay of the phase information can be written as a function of the traveled distance L, which is the circumference of the ring, according to

$$\frac{1}{2}\langle \delta\varphi^2 \rangle = \frac{L}{l_\varphi(T)}.$$

If we consider decoherence in the expression for the conductance of the quantum ring, each interference term, i.e., (14.9) and (14.10), obtains a new prefactor which dampens the oscillation amplitude. For example, the h/e-periodic Aharonov–Bohm oscillations decay according to

$$\int_{-\infty}^{+\infty} d\varphi \frac{1}{\sqrt{2\pi\langle \delta\varphi^2 \rangle}} \exp\left(-\frac{(\varphi-\delta)^2}{2\langle \delta\varphi^2 \rangle}\right) \cos\varphi = e^{-L/l_\varphi(T)} \cos\delta.$$

The complete result for the reflection probability is

$$\begin{aligned} \mathcal{R} \;=\; & r_0^2 + 2r_1^2 + \ldots \\ & + 4\,|r_0|\,|r_1|\, e^{-L/l_\varphi(T)} \cos\delta \cos\left(2\pi\frac{\phi}{\phi_0}\right) + \ldots \\ & + 2\,|r_1|^2\, e^{-2L/l_\varphi(T)} \cos\left(4\pi\frac{\phi}{\phi_0}\right) + \ldots \end{aligned}$$

Figure 14.11 shows the amplitude of h/e- and $h/2e$-periodic oscillations as a function of temperature. The logarithmic plot makes clear that in this experiment the amplitudes are indeed proportional to $e^{-\alpha T}$, where α is a constant depending on the oscillation period, as expected for decoherence from coupling to a thermal bath. Because the $h/2e$-periodic

oscillations have an effective path length twice as long as those with the h/e period, we expect $\alpha_{h/e} = \alpha_{h/(2e)}/2$. However, the h/e-periodic oscillations are strongly affected by averaging over the energy (derivative of the Fermi function in the expression for the conductance), whereas the $h/2e$ oscillations are not. Therefore, in the experiment $\alpha_{h/e} > \alpha_{h/(2e)}/2$. The h/e-periodic Aharonov–Bohm oscillations are therefore not so well suited for a determination of the phase-coherence time, and it is preferable to use the decay of the $h/2e$-periodic Altshuler–Aronov–Spivak oscillations for the determination of τ_φ. In our example, we find

$$l_\varphi(T) \propto T^{-1} \Rightarrow \frac{1}{\tau_\varphi} \propto T,$$

i.e., the decoherence rate increases linearly with temperature as expected for a thermal bath.

14.8 Conductance fluctuations in mesoscopic samples

The Aharonov–Bohm effect is the basis for the understanding of magnetoconductance fluctuations occurring also in singly connected mesoscopic samples. This phenomenon occurs in ballistic systems, where the elastic mean free path l_e is larger than the system size L, i.e., $l_e \gg L$, or in diffusive systems with $l_e \ll L$. In both cases, the phenomenon arises only if the phase-coherence length l_φ is larger than the system size L, i.e., if $l_\varphi \gg L$. Typical values $l_\varphi > 1\,\mu\mathrm{m}$ can be reached in experiments below 1 K in high quality samples.

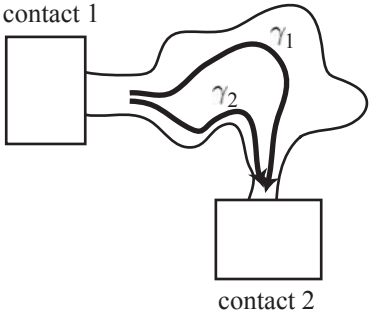


Fig. 14.24 Mesoscopic sample with two contacts. The paths γ_1 and γ_2 are two of many paths contributing to the total transmission from contact 1 to contact 2, and thereby to the conductance.

Ballistic conductance fluctuations. We start the discussion with the ballistic variant. Fig. 14.24 shows schematically a mesoscopic sample with two contacts. The conductance of the sample is given by the transmission between the two contacts. According to the rules of quantum mechanics (Feynman *et al.*, 2006) the transmission probability is the square of the sum of many complex-valued transmission amplitudes of different paths between the two contacts. Two such paths, γ_1 and γ_2 are indicated schematically in the figure. Many pairs of these paths have a common starting and end point. They therefore enclose a certain area. In a small magnetic field (we again require $R_\mathrm{c} \gg L$), such pairs lead to interference terms in the total transmission having an Aharonov–Bohm period corresponding to the enclosed area. Therefore the conductance in a magnetic field contains an interference contribution

$$G_\mathrm{int} = \sum_{mn} |t_m|\,|t_n| \cos(\theta_m - \theta_n) \cos(2\pi e B A_{mn}/h),$$

where A_{mn} is the area enclosed by the paths m and n. Although the statistical distribution of the areas A_{mn} and of the phase differences $\theta_m - \theta_n$ leads to averaging where interference contributions cancel each other, averaging is not complete in a sufficiently small sample. Even

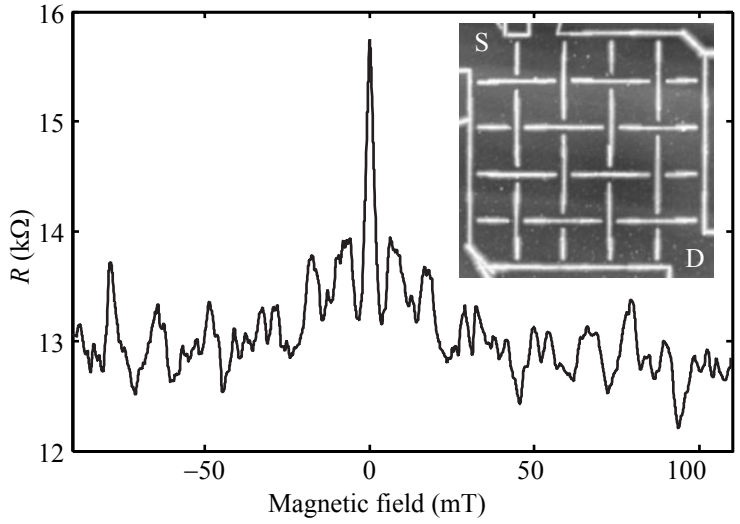


Fig. 14.25 Measurement of ballistic conductance fluctuations in a sample consisting of 25 mesoscopic squares, coupled via quantum point contacts. Current is driven from source (S) to drain (D). The sample was patterned using AFM lithography on a two-dimensional electron gas in the Ga[Al]As material system.

at zero magnetic field, fluctuations can be observed as a function of the electron density in the structure, because the phase differences of paths depend on the Fermi velocity, which can usually be tuned using gate voltages. These *ballistic conductance fluctuations (BCF)* disappear with increasing temperature as a result of thermal averaging and decoherence, in complete analogy to the disappearance of Aharonov–Bohm oscillations in ring structures. The ballistic effect can be observed if the elastic mean free path l_e is large compared to the sample size L as long as the sample size L is not much larger than the phase coherence length l_φ.

Figure 14.25 shows an example for ballistic conductance fluctuations in the sample that is shown as an inset in the upper right corner. The source contact is at the top left, the drain contact at the bottom right. Electrons move ballistically in the $1\,\mu\mathrm{m} \times 1\,\mu\mathrm{m}$ squares which are connected by narrow constrictions. The fluctuations appear as a function of magnetic field and are completely stable and reproducible in time.

Universal conductance fluctuations. The same effect also arises in diffusive systems, where the electronic motion is strongly influenced by elastic scattering. In such systems, however, the relation between the phase coherence length l_φ and the coherence time τ_φ is no longer given by eq. (14.31), but by the relation

$$l_\varphi = \sqrt{D\tau_\varphi}.$$

Here, $D = v_\mathrm{F} l_\mathrm{e}/2$ is the diffusion constant for two-dimensional systems. In diffusive systems, the effect is known as *universal conductance fluctuations (UCF)*, because it is theoretically predicted that for temperatures $T \to 0$ the amplitude of the fluctuations is of the order of the conductance quantum e^2/h, if the phase-coherence length l_φ is larger, or of

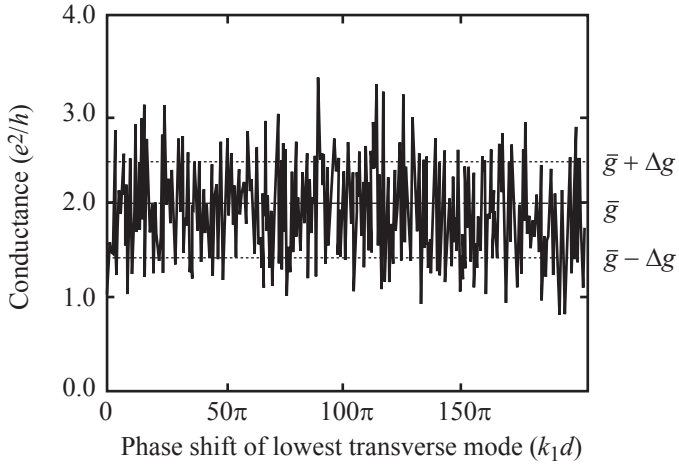


Fig. 14.26 Calculated conductance of a diffusive quantum wire containing a random array of 600 scatterers, as the position of a single impurity in the middle of the wire is changed. The fluctuations of the conductance Δg are of the order of e^2/h, $\bar{g}$ is the average conductance. (Reprinted with permission from Cahay *et al.*, 1988. Copyright 1988 by the American Physical Society.)

the order of the system size (Altshuler, 1985; Lee and Stone, 1985; Lee, 1986). This fluctuation amplitude is independent of the sample size and the strength of the disorder, a fact that made people call the fluctuations *universal*.

In diffusive systems, the fluctuations arise if the impurity configuration in the sample is changed. Figure 14.26 shows the calculated conductance of a fully coherent quantum wire with 30 modes in which 600 scattering sites have been distributed randomly (Cahay *et al.*, 1988). Along the horizontal axis, a single impurity in the middle of the wire is shifted. As a result, the calculated conductance is seen to fluctuate with an amplitude of the order of e^2/h. The characteristic length scale over which a shift causes a conductance change of this order of magnitude is the Fermi wavelength of the electrons.

The magnitude of the fluctuations at zero temperature can be understood with the following argument (Lee, 1986): the conductance of a two-terminal device is in general given by eq. (11.13) which has the zero temperature limit

$$G = \frac{e^2}{h} \sum_{n,m} \mathcal{T}_{nm}(E_\mathrm{F}) = \frac{e^2}{h} \left[N - \sum_{n,m} \mathcal{R}_{nm}(E_\mathrm{F}) \right],$$

where the sum is extended over the number N of occupied modes (here we assume that each spin orientation counts as one mode). The average conductance is then

$$\langle G \rangle = \frac{e^2}{h} \left[N - \left\langle \sum_{n,m} \mathcal{R}_{nm}(E_\mathrm{F}) \right\rangle \right] = \frac{e^2}{h} N \left[1 - N \langle \mathcal{R}_{nm}(E_\mathrm{F}) \rangle \right]. \quad (14.32)$$

The magnitude of fluctuations due to varying disorder configuration is

$$\mathrm{Var} G = \langle (G - \langle G \rangle)^2 \rangle = \left(\frac{e^2}{h} \right)^2 \mathrm{Var} \left[\sum_{n,m} \mathcal{R}_{nm}(E_\mathrm{F}) \right].$$

Here we have to determine the variance of the sum of N^2 quantities $\mathcal{R}_{nm}$. If we assume that these quantities are uncorrelated, then, according to the central limit theorem,

$$\mathrm{Var}G = \left(\frac{e^2}{h}\right)^2 N^2 \mathrm{Var}\left[\mathcal{R}_{nm}(E_\mathrm{F})\right].$$

The variance of the reflection probabilities $\mathcal{R}_{nm} = |r_{nm}|^2$ is now determined in the following way: each reflection amplitude r_{nm} is composed of a large number of different probability amplitude contributions A_i (i is an integer index) involving distinct paths or scattering sequences within the sample. An example of two such paths A_1 and A_2 is shown in Fig. 14.27. More specifically, this leads to

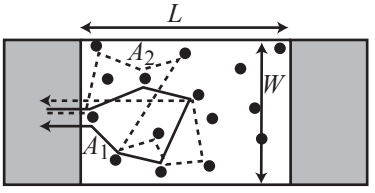


Fig. 14.27 Schematic sketch of a strongly disordered sample of width W and length L. Two specific paths contributing to the reflection probability with probability amplitudes A_1 and A_2 are shown as solid and dashed lines. They differ in their scattering sequences.

$$\mathcal{R}_{nm} = \left|\sum_i A_i\right|^2 = \sum_{ij} A_i A_j^\star.$$

In order to calculate the required variance we need $\langle \mathcal{R}_{nm}^2 \rangle$ and $\langle \mathcal{R}_{nm} \rangle^2$. The former is found from

$$\langle \mathcal{R}_{nm}^2 \rangle = \sum_{ijkl} \langle A_i A_j^\star A_k A_l^\star \rangle.$$

We now assume the phases of different paths to be completely random, such that most paths cancel each other. However, pairs of paths with $i = j$ and $k = l$, or $i = l$ and $j = k$ give the contribution

$$\langle \mathcal{R}_{nm}^2 \rangle = \sum_{ijkl} \langle |A_i|^2 \rangle \langle |A_k|^2 \rangle (\delta_{ij}\delta_{kl} + \delta_{il}\delta_{jk}) = 2 \sum_{ik} \langle |A_i|^2 \rangle \langle |A_k|^2 \rangle.$$

On the other hand, based on the same random phase argument we find

$$\langle \mathcal{R}_{nm} \rangle^2 = \sum_{ik} \langle |A_i|^2 \rangle \langle |A_k|^2 \rangle = \frac{1}{2} \langle \mathcal{R}_{nm}^2 \rangle,$$

and therefore obtain

$$\mathrm{Var}G = \left(\frac{e^2}{h}\right)^2 N^2 \langle \mathcal{R}_{nm} \rangle^2. \tag{14.33}$$

The remaining problem is therefore to estimate the mean reflection $\langle \mathcal{R}_{nm} \rangle$. We do this using the Drude result for the conductance which is

$$\langle G \rangle = \frac{W}{L} \frac{n_\mathrm{s} e^2 \tau}{m} = \frac{W}{L} \frac{e^2}{h} k_\mathrm{F} l_\mathrm{e} = \frac{e^2}{h} \frac{W}{\lambda_\mathrm{F}/2} \frac{\pi l_\mathrm{e}}{L}.$$

If we now realize that in a channel of width W the second fraction is about the number of modes N, we have

$$\langle G \rangle = \frac{e^2}{h} N \frac{\pi l_\mathrm{e}}{L}.$$

Comparing this expression with eq. (14.32), we find the estimate

$$\langle \mathcal{R}_{nm} \rangle = \frac{1}{N}\left(1 - \frac{\pi l_{\rm e}}{L}\right) \approx \frac{1}{N},$$

because $l_{\rm e}/L \ll 1$ in a diffusive sample. Inserting this estimate into eq. (14.33) leads to

$$\mathrm{Var}G \approx \left(\frac{e^2}{h}\right)^2, \quad \text{or} \quad \Delta G = \sqrt{\mathrm{Var}G} \approx \frac{e^2}{h}.$$

The relative magnitude of the fluctuations is

$$\frac{\Delta G}{\langle G \rangle} = \frac{\Delta R}{\langle R \rangle} \approx \frac{1}{N}\frac{L}{\pi l_{\rm e}} = \frac{L}{W}\frac{1}{k_{\rm F} l_{\rm e}},$$

implying that the relative magnitude of the fluctuations increases with the length of the channel, i.e., no self-averaging occurs as long as $L \leq l_\varphi$. In high mobility samples with $k_{\rm F} l_{\rm e} \gg 1$ the oscillations will be very small, whereas they appear to be strong in lower mobility samples.

The exact result of Lee and Stone, 1985, for the magnitude of the fluctuations is

$$\Delta G = \frac{g_{\rm s} g_{\rm v}}{2} \beta^{-1/2} C \frac{e^2}{h},$$

where C is a geometry-dependent constant of order unity, $\beta = 1$ at zero magnetic field and $\beta = 2$ at a finite magnetic field breaking time-reversal symmetry, and the factors $g_{\rm s}$ and $g_{\rm v}$ are spin and valley degeneracy factors, respectively. A table with relevant results for C can be found in Beenakker and van Houten, 1991.

At finite temperature, the conductance fluctuations are reduced in amplitude for two reasons: first, the finite phase-coherence length $l_\varphi(T)$ leads to averaging of independently fluctuating segments of a sample, and second, the smearing of the Fermi distribution function results in energy averaging. Let us, for example, consider the effect of the finite-phase coherence length in a case where $W < l_\varphi < L$. Then we can consider L/L_φ segments of the channel fluctuating independently. If the average resistance of one segment of length l_φ is R_0, then we find

$$\Delta G = \frac{\Delta R}{\langle R \rangle^2} = \frac{\Delta R_0 \sqrt{L/l_\varphi}}{\langle R_0 \rangle^2 (L/l_\varphi)^2} \approx \frac{e^2}{h}\left(\frac{l_\varphi}{L}\right)^{3/2}.$$

In order to take energy averaging at finite temperature into account we define a correlation energy $E_{\rm c}$ related to the phase coherence length l_φ. If we consider two paths propagating at energies differing by $E_{\rm c}$ as uncorrelated if their phase difference $\Delta\varphi = E_{\rm c}\tau_\varphi/\hbar \approx 1$, we are led to the correlation energy

$$E_{\rm c} = \frac{\hbar D}{l_\varphi^2}.$$

Comparing this correlation energy to the thermal broadening of the Fermi distribution function gives the number of independently fluctuating energy channels that contribute to the conductance. In contrast to

the previous argument, where independently fluctuating segments were connected in series, here the channels are contributing in parallel. The fluctuation amplitude is therefore further reduced and becomes

$$\Delta G \approx \frac{e^2}{h}\left(\frac{l_\varphi}{L}\right)^{3/2}\left(\frac{k_\mathrm{B}T}{E_\mathrm{c}}\right)^{-1/2} = \frac{e^2}{h}\left(\frac{l_\varphi}{L}\right)^{3/2}\frac{l_T}{l_\varphi} \quad \text{for } l_\varphi \gg l_\mathrm{T},$$

where we have introduced the thermal length $l_T^2 = \hbar D/k_\mathrm{B}T$. We see here that increasing temperature will reduce the amplitude of the fluctuations. However, it also becomes clear that the phase-coherence length l_φ can be estimated from the magnitude of the fluctuations. A very useful interpolation formula given in Beenakker and van Houten, 1988*a*, is

$$\Delta G = \frac{g_\mathrm{s}g_\mathrm{v}}{2}\beta^{-1/2}\sqrt{12}\frac{e^2}{h}\left(\frac{l_\varphi}{L}\right)^{3/2}\left[1+\frac{9}{2\pi}\left(\frac{l_\varphi}{l_T}\right)^2\right]^{-1/2}.$$

As in the ballistic case, the fluctuations of the conductance can be studied experimentally as a function of magnetic field, and similar fluctuations will be found on a characteristic correlation field scale ΔB_c. This field scale can be determined from measured data from the full width at half maximum of the autocorrelation function of the fluctuations. The theory for conductance fluctuations in a magnetic field has been worked out by Lee *et al.*, 1987. The basic idea is that a change of the magnetic field by ΔB_c is equivalent to the measurement of a sample with a different impurity configuration. It is found that (Lee *et al.*, 1987)

$$\Delta B_\mathrm{c} = C\frac{\Phi_0}{W l_\varphi},$$

where $\Phi_0 = h/e$ is the magnetic flux quantum, and C is a prefactor of order unity which increases from 0.42 for $l_\varphi \ll l_T$ to 0.95 for $l_\varphi \gg l_T$ (Beenakker and van Houten, 1988*a*). The correlation field provides another experimental way of estimating the phase-coherence length l_φ.

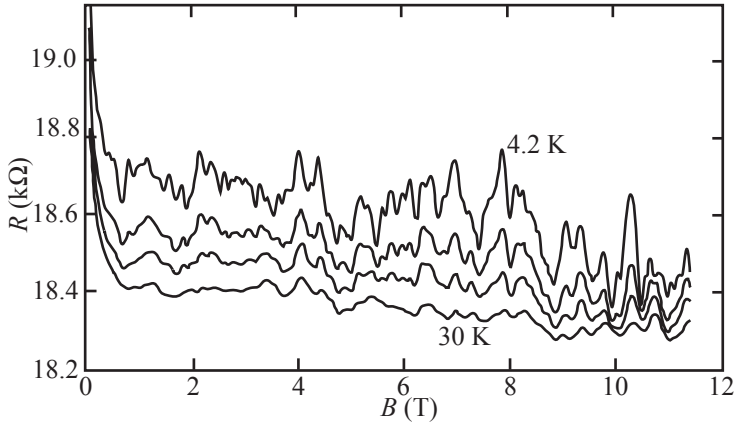


Fig. 14.28 Measured magnetoresistance of a 260 nm wide heavily doped quantum wire in GaAs at temperatures $T = 4.2\,\mathrm{K}$, $11.8\,\mathrm{K}$, $17.7\,\mathrm{K}$, and $30\,\mathrm{K}$. (Reprinted from Taylor *et al.*, 1988 with permission from Elsevier.)

Figure 14.28 shows an example of the magnetoresistance measured on a strongly disordered quantum wire in GaAs with $W = 260\,\mathrm{nm}$. The fluctuations become weaker with increasing temperature, but some persist up to $T = 30\,\mathrm{K}$.

Classical conductance fluctuations. Fluctuations of the classical conductance are in most cases negligibly small. In this context, classical means that the phase-coherence length l_φ is smaller than the elastic mean free path l_e. For example, in a narrow wire of length L, the conductance may then be seen as the incoherent series addition of the large number L/l_e of independently fluctuating segments. Classical fluctuations in the resistance will then, according to the central limit theorem, scale with the wire length according to $(l_\mathrm{e}/L)^{1/2}$.

Further reading

- Aharonov–Bohm effect: Beenakker and van Houten 1991; Datta 1997; Imry 2002.
- Decoherence: Imry 2002.
- Interferometry with electrons in semiconductor nanostructures: Gefen 2002.
- Probability amplitudes in quantum mechanics: Feynman *et al.* 2006
- Papers: Aharonov and Bohm 1959; Berry 1984; Stern *et al.* 1990.

Exercises

(14.1) Diffusive electron motion in the quantum regime, where the phase-coherence length l_φ is much larger than the elastic mean free path l_e, can be described successfully within a semiclassical approach if the Fermi wavelength λ_F is much smaller than l_e. Within this approach, the classical trajectories of single electrons are determined first. In a second step, one considers wave propagation along these classical trajectories which adds the quantum aspect to the problem and allows for interference. In this spirit we consider the two-dimensional diffusive motion of classical electrons in more detail. It is governed by the diffusion equation

$$\partial_t C(\mathbf{r},t) = D\Delta C(\mathbf{r},t),$$

with the diffusion constant D. The quantity $C(\mathbf{r},t)d\mathbf{r}$ is the probability that the diffusing particle is found in a small region $d\mathbf{r}$ of space around the position $\mathbf{r}$, after travelling for a time t.

(a) Suppose that, at $t = 0$, the particle starts at $\mathbf{r} = 0$, i.e., $C(\mathbf{r},0) = \delta(\mathbf{r})$. Solve the diffusion equation for this initial condition. Hint: Transform the equation from real space into Fourier space and then solve the time-dependent first-order linear differential equation.

(b) Now consider only trajectories that start at the origin and end up at a particular location A, a distance L from the origin. What is the distribution of travel times for these trajectories? Show that the most likely time that an electron has spent diffusing, before it reached A, is given by $t_\mathrm{diff} = L^2/4D$.

(c) Calculate the most likely trajectory length L_{diff} of electrons diffusing from the origin to A. How does L_{diff} compare to the length L of the ballistic trajectory? Hint: physical meaning is achieved by expressing the diffusion constant in terms of the elastic mean free path l_{e}.

(d) Now we make the transition from particle diffusion to waves. What is the most likely phase φ_{diff} that an electron wave acquires along the classical diffusive path from the origin to A? (Neglect possible phase shifts due to the scattering events.) Show that φ_{diff} can be expressed as

$$\varphi_{\mathrm{diff}} = \frac{1}{2}\frac{E_{\mathrm{F}}}{E_{\mathrm{Th}}},$$

and determine an expression for the energy scale E_{Th}. This energy scale is called the Thouless energy (you will learn more about it in the next chapter).

(e) Assume that E_{Th} does not change if the Fermi energy changes by a small amount. How much does the energy of an electron near the Fermi energy need to change, such that φ_{diff} changes by $\pi/2$? Discuss, in what respect E_{Th} can be called a (phase) correlation energy.

This page intentionally left blank

Diffusive quantum transport

15

15.1 Weak localization effect

In the previous chapter we looked into the physics of electron interference. Initially we considered idealized ballistic ring geometries, but later we extended our considerations to conductance fluctuations in diffusive systems. Most importantly, the phase-coherence length l_φ entered our physical picture of semiconductor nanostructures, and we discussed the basics of decoherence. Diffusive systems were found to exhibit manifestations of electron interference in the mesoscopic regime, i.e., if the sample size L was not much larger than l_φ. In this chapter we are going to extend the discussion of interference effects to large diffusive systems, where $L \gg l_\varphi$.

Within the Drude model, quantum mechanical scattering at individual impurities is considered, but the coherent motion between scattering events and multiple scattering are neglected. In the 1980s scattering theories were developed in which phase-coherent multiple scattering was taken into account systematically. It was found that these processes led to enhanced backscattering of electrons and thereby to a logarithmic increase of the resistance. This effect is called *weak localization* and can be traced back to the constructive interference of time-reversed Feynman paths of electrons (Bergmann, 1983).

Figure 15.1 shows an example of two time-reversed paths. Each of the two partial waves returns after multiple scattering to the starting point, but the loop is traveled by them in opposite directions. If we denote the complex quantum mechanical amplitudes of the two paths by A^+ and A^-, the probability of return to the starting point is given by

$$\left|A^+ + A^-\right|^2 = \left|A^+\right|^2 + \left|A^-\right|^2 + A^+ A^{-\star} + A^{+\star} A^-.$$

Here, the first two terms on the right-hand side are the so-called classical contributions to backscattering which are contained in the Drude–Boltzmann theory. The last two terms are interferences which are neglected in the incoherent approximation of Drude–Boltzmann, but of crucial importance for the weak localization correction. It is important to note that, at zero magnetic field, time-reversal symmetry requires $A^+ = A^- \equiv A$. The classical contribution to the return probability is then given by

$$P_{\mathrm{cl}} = 2\left|A\right|^2,$$

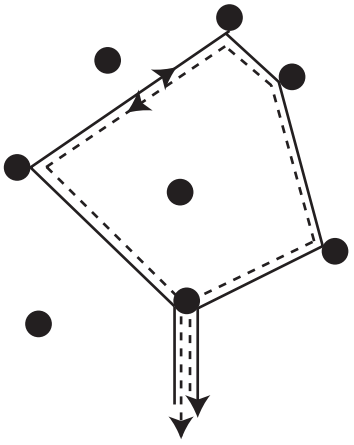


Fig. 15.1 Time-reversed paths in a diffusive two-dimensional sample. At zero magnetic field the interference of such paths is always constructive.

whereas the quantum mechanical return probability is enhanced by a factor of two as a result of the interference term, and we obtain

$$P_{\rm qm} = 4\,|A|^2\,.$$

In this sense we talk about enhanced backscattering, and of weak localization, as a result of quantum interference.

We have seen that the interference correction to the classical return probability is equal to the classical return probability. Therefore we can semiquantitatively describe the weak localization correction to the conductivity at zero magnetic field by considering the classical diffusion of electrons (Khmel'nitskii, 1984). Here we restrict the discussion to the two-dimensional case. Let $C(\mathbf{r})d^2r$ be the classical probability that a diffusing particle returns to the volume element d^2r around its starting point after time t. From the solution of the classical diffusion equation we find

$$C(t) = \frac{1}{4\pi D t}$$

for time scales on which the electron has travelled distances larger than an elastic mean free path. If we form semiclassical wave packets from the occupied electronic states in the system, we cannot localize an electron better than on the scale of $\lambda_{\rm F}$. We are therefore interested in the return probability into the area $\lambda_{\rm F}^2 = v_{\rm F}\Delta t\lambda_{\rm F} = h\Delta t/m^\star$ which is given by

$$\frac{h}{m^\star}\frac{\Delta t}{4\pi D t}.$$

In order to obtain the total classical return probability for travel times between $\tau_{\rm e}$ and τ_φ (longer diffusive paths cannot interfere) we have to sum this expression over time intervals Δt. Performing an integral instead of the sum we obtain the classical return probability

$$p_{\rm ret} = \frac{h}{m^\star}\int_{\tau_{\rm e}}^{\tau_\varphi}\frac{dt}{4\pi D t} = \frac{\hbar}{2m^\star D}\ln\frac{\tau_\varphi}{\tau_{\rm e}}.$$

As a brief aside we can show that cutting the time integral more smoothly gives essentially the same result, as long as $\tau_\varphi \gg \tau_{\rm e}$. For example,

$$p_{\rm ret} = \frac{h}{m^\star}\int_0^\infty \frac{dt}{4\pi D t}(1-e^{-t/\tau_{\rm e}})e^{-t/\tau_\varphi} = \frac{\hbar}{2m^\star D}\ln\left(1+\frac{\tau_\varphi}{\tau_{\rm e}}\right) \approx \frac{\hbar}{2m^\star D}\ln\frac{\tau_\varphi}{\tau_{\rm e}}.$$

Taking into account that in two dimensions $D = v_{\rm F}^2\tau_{\rm e}/2$, the prefactor can be rewritten as $1/k_{\rm F}l_{\rm e}$. The normalized quantum correction of the Drude conductivity is proportional to $p_{\rm ret}$, i.e.,

$$\frac{\delta\sigma_{\rm qm}}{\sigma} \approx -p_{\rm ret} = -\frac{1}{k_{\rm F}l_{\rm e}}\ln\frac{\tau_\varphi}{\tau_{\rm e}}. \tag{15.1}$$

This is a logarithmic quantum correction to the classical Drude conductivity, which is negative, i.e., it reduces the conductivity, as a result of the weak localization caused by the constructive interference of time-reversed paths.

15.2 Decoherence in two dimensions at low temperatures

We discussed the general ideas behind decoherence in section 14.7. Here we are specifically interested in decoherence mechanisms in two-dimensional electron gases at low temperatures. A two-dimensional electron gas exists in a crystal lattice in which lattice vibrations can be excited. The interaction of electrons and phonons are inelastic and impair the phase coherence of the electrons. Experiments aimed at quantum properties are typically performed at low temperatures below 4.2 K (liquid He), because lattice vibrations are frozen out. At such low temperatures electron–electron interactions are found to dominate the decoherence of electron waves. A single electron is surrounded by a sea of other electrons. The random motion of the latter creates an electromagnetic field randomly fluctuating in time (photons) which can scatter the electron. Elastic scattering processes, such as, for example, scattering at static spatial potential fluctuations conserve the phase coherence of the electrons.

Decoherence is described using the decoherence rate $1/\tau_\varphi$. It usually depends on temperature T and often follows a power law,

$$\frac{\hbar}{\tau_\varphi} \propto (k_\mathrm{B}T)^p.$$

For the particular decoherence rate of electrons in two-dimensional electron gases at low temperatures, a number of authors (Altshuler and Aronov, 1985; Chakravarty and Schmid, 1986; Imry, 2002) have found the equation

$$\frac{\hbar}{\tau_\varphi} = k_\mathrm{B}T\frac{e^2/\hbar}{\sigma}\ln\frac{k_\mathrm{B}T}{\hbar/\tau_\varphi} \overset{\sigma\gg e^2/\hbar}{\approx} k_\mathrm{B}T\frac{e^2/\hbar}{\sigma}\ln\frac{\sigma}{e^2/\hbar}. \tag{15.2}$$

It has the general form

$$x = -\frac{1}{a}\ln x$$

with $x = \hbar/\tau_\varphi k_\mathrm{B}T$ and $a = \sigma/(e^2/\hbar)$, with the solution $x = f(a)$ which has to be determined numerically. The solution of this equation is shown in Fig. 15.2. The decoherence rate can then be written as

$$\frac{\hbar}{\tau_\varphi} = k_\mathrm{B}T\, f\left(\frac{\sigma}{e^2/\hbar}\right). \tag{15.3}$$

It can be seen that at a given temperature T the decoherence rate decreases with increasing conductivity σ. For a GaAs heterostructure with an electron density $n = 3\times10^{15}\,\mathrm{m}^{-2}$ and mobility $\mu = 10^6\,\mathrm{cm}^2/\mathrm{Vs}$ we find $\sigma/(e^2/\hbar) = 197$. From Fig. 15.2 we read $f(197) \approx 0.02$. At a typical temperature of a ^{3}He cryostat of 300 mK this gives a coherence time $\tau_\varphi = 1.3\,\mathrm{ns}$, compared to an elastic scattering time $\tau_\mathrm{e} = 38\,\mathrm{ps}$. For the same density, but a mobility of $\mu = 5\times10^4\,\mathrm{cm}^2/\mathrm{Vs}$ we find $f(a) = 0.2$ and therefore at 300 mK $\tau_\varphi = 127\,\mathrm{ps}$ compared to $\tau_\mathrm{e} = 2\,\mathrm{ps}$.

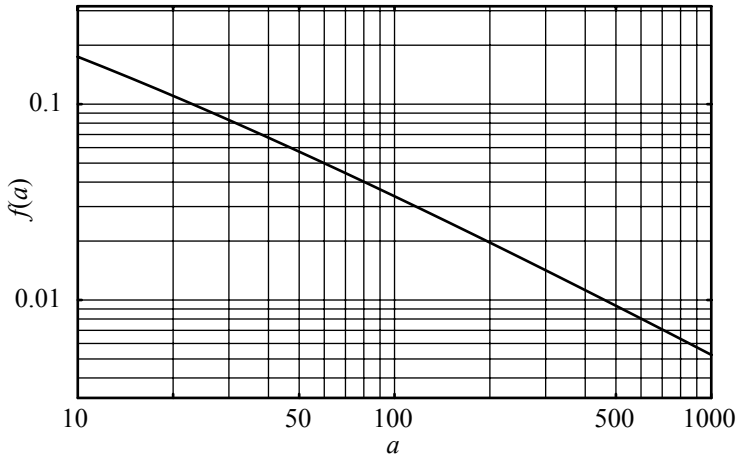


Fig. 15.2 The function $f(a)$. It determines the ratio of $\hbar/\tau_\varphi$ and $k_\mathrm{B}T$, if the parameter a, given by $\sigma/(e^2/\hbar)$, is known.

The phase coherence time is much larger than the Drude scattering time at sufficiently low temperatures.

During the phase coherence time τ_φ electrons therefore move diffusively and we can define the phase coherence length

$$l_\varphi = \sqrt{D_d \tau_\varphi},$$

where D_d is the diffusion constant in d dimensions. For the two above examples we find $l_\varphi = 37\,\mu\mathrm{m}$ and $l_\varphi = 2.6\,\mu\mathrm{m}$, respectively.

15.3 Temperature-dependence of the conductivity

Experimentally the weak localization effect in the absence of a magnetic field can be measured via the temperature dependence of the conductance or the resistance of a sample, because the phase coherence length l_φ depends on temperature. Nevertheless we have to take into account that both the Drude conductivity and interaction effects contribute with their own temperature dependence. However, these effects are relatively weak in thin metallic films.

Figure 15.3 shows the measured temperature-dependent resistance of a AuPd film with a thickness of only a few nanometers. The resistance increases logarithmically with decreasing temperature. This behavior can be explained with the help of eqs. (15.1) and (15.3). Combining the two gives

$$\frac{\delta\sigma}{\sigma} = \frac{1}{k_\mathrm{F} l_\mathrm{e}} \ln\left(k_\mathrm{B}T \frac{f(k_\mathrm{F} l_\mathrm{e}/2\pi)}{\hbar/\tau_\mathrm{e}}\right).$$

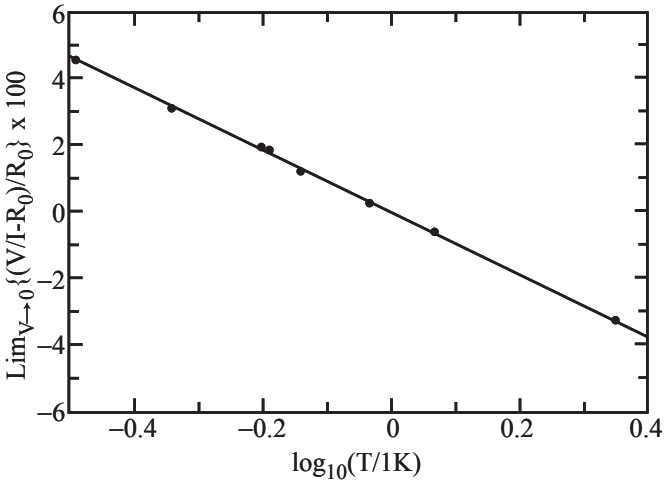


Fig. 15.3 Temperature dependence of the resistance of a thin metal film between 300 mK and 2.5 K. The resistance increases logarithmically with decreasing temperature. (Reprinted with permission from Dolan and Osheroff, 1979. Copyright 1979 by the American Physical Society.)

15.4 Suppression of weak localization in a magnetic field

The microscopic picture of phase coherent backscattering leading to weak localization is a good starting point for understanding the magnetoconductivity. The magnetic field influences the phases of the amplitudes A^+ and A^- by adding an Aharonov–Bohm phase. The Aharonov–Bohm phase and its effects in ring-shaped nanostructures were discussed in chapter 14. Here we recall that the magnetic field acts on the phases of the amplitudes such that

$$A^{\pm}(B) = Ae^{\pm i\varphi_{\mathrm{AB}}},$$

[cf. eqs (14.2) and (14.3)], where the Aharonov–Bohm phase φ_{AB} is given by

$$\varphi_{\mathrm{AB}} = 2\pi\frac{BS}{h/e},$$

with S being the area enclosed by the two counterpropagating paths. This leads to

$$\left|A^+(B) + A^-(B)\right|^2 = 2\left|A\right|^2 + 2\left|A\right|^2\cos\left(4\pi\frac{BS}{h/e}\right).$$

The magnetic field leads to an $h/2e$-periodic modulation of the quantum interference correction of the backscattering probability. The interference term describes an effect related to the $h/2e$-periodic Altshuler–Aronov–Spivak oscillations which were discussed in chapter 14. In a macroscopic diffusive sample many pairs of time-reversed paths enclosing different areas will occur and we may introduce a probability density distribution $P(S)dS$ denoting the relative contribution of areas of size S to the total conductivity. As a result we can expect that the oscillatory contribution of individual time-reversed paths averages out completely

at finite magnetic fields. Close to zero field, however, all these oscillations have the same phase (maximum backscattering, the cosine has its maximum at $B = 0$ for all paths) and a minimum in the conductivity survives the averaging procedure. The result is a magnetic-field-dependent quantum correction $\delta\sigma(B)$ of the conductance which has the form

$$\frac{\delta\sigma_{\rm qm}(B)}{\sigma} = -\frac{1}{k_{\rm F} l_{\rm e}} \int dS P(S, l_\varphi) \cos\left(4\pi \frac{BS}{h/e}\right). \tag{15.4}$$

This is the Fourier cosine transform of the function $P(S, l_\varphi)$. It leads to a minimum of the conductance at $B = 0$ in macroscopic samples. The expression of $\delta\sigma_{\rm qm}/\sigma$ at zero magnetic field is exactly that given in eq. (15.1).

The smallest areas S contributing to backscattering are of the order of $l_{\rm e}^2$, whereas the largest are of the order l_φ^2. The sharpness (i.e., the curvature) of the conductance minimum at $B = 0$ will be determined by the phase coherence length l_φ, because the largest areas lead to the contributions with the smallest periods. Looking at eq. (15.1) we may identify the quantity Dt with an effective area S and write heuristically for the probability distribution

$$P(S, l_\varphi) dS \propto \frac{1}{S} \left(1 - e^{-S/l_{\rm e}^2}\right) e^{-S/l_\varphi^2} dS. \tag{15.5}$$

Inserting this expression in eq. (15.4) and performing the integration over areas leads to typical magnetoresistance corrections of the shape shown in Fig. 15.4 (note that $\delta\rho/\rho = -\delta\sigma/\sigma$).

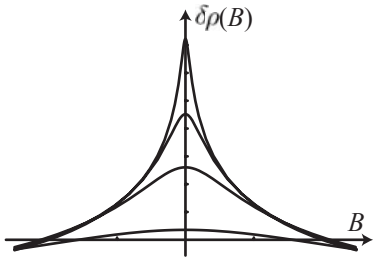


Fig. 15.4 Magnetoresistance correction calculated from eq. (15.4) for different ratios $l_\varphi/l_{\rm e}$. Large values of this ratio lead to sharper maxima at zero field.

The accurate quantitative theory of the weak localization effect has been worked out using diagrammatic methods (Mahan, 2000) that are beyond the scope of this book. The result for the magnetic-field-dependent correction of the Drude conductance is

$$\delta\sigma(B) - \delta\sigma(0) = \frac{e^2}{2\pi^2\hbar} \left[\Psi\left(\frac{1}{2} + \frac{\tau_B}{2\tau_\varphi}\right) - \Psi\left(\frac{1}{2} + \frac{\tau_B}{2\tau_{\rm e}}\right) + \ln\left(\frac{\tau_\varphi}{\tau_{\rm e}}\right)\right]. \tag{15.6}$$

Here, $\tau_B = \hbar/(2eDB)$ and $\Psi(x)$ is the digamma function. This relation is correct for $W, L \gg \tau_\varphi \gg \tau_{\rm e}$.

Figure 15.5 shows a measurement of the effect as it is seen in a p-doped SiGe quantum well structure. The reduction of the conductivity at zero magnetic field corresponds to an enhancement of the resistivity. With increasing magnetic field the weak localization effect is suppressed. Beyond a magnetic field scale of $B = h/(2el_{\rm e}^2)$ it has disappeared completely. At elevated magnetic fields and low temperatures, Shubnikov–de Haas oscillations are visible (see chapter 16). With increasing temperature the width of the peak increases and its height decreases, because the ratio $l_\varphi/l_{\rm e}$ decreases.

The weak localization effect disappears with increasing temperature, because l_φ becomes smaller than $l_{\rm e}$. The quantum correction is quenched as soon as $l_\varphi \approx l_{\rm e}$. In samples of very high mobility, $l_\varphi < l_{\rm e}$ at all temperatures that are accessible by experiment (i.e., down to about 5 mK),

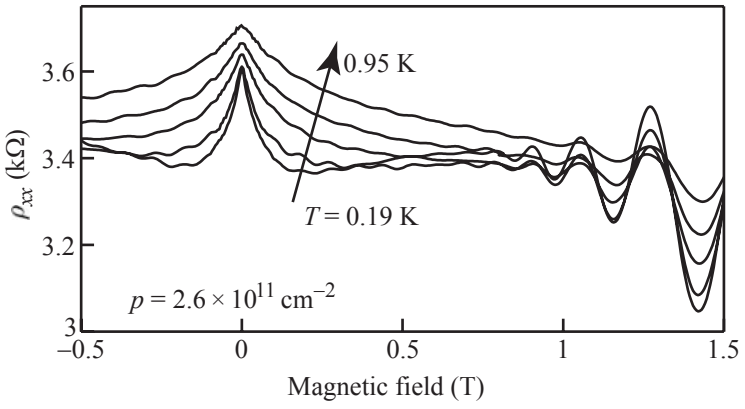


Fig. 15.5 Weak localization effect at different temperatures, measured in a two-dimensional hole gas residing in a SiGe quantum well. (Reprinted with permission from Senz *et al.*, 2000*a*. Copyright 2000 by the American Physical Society.)

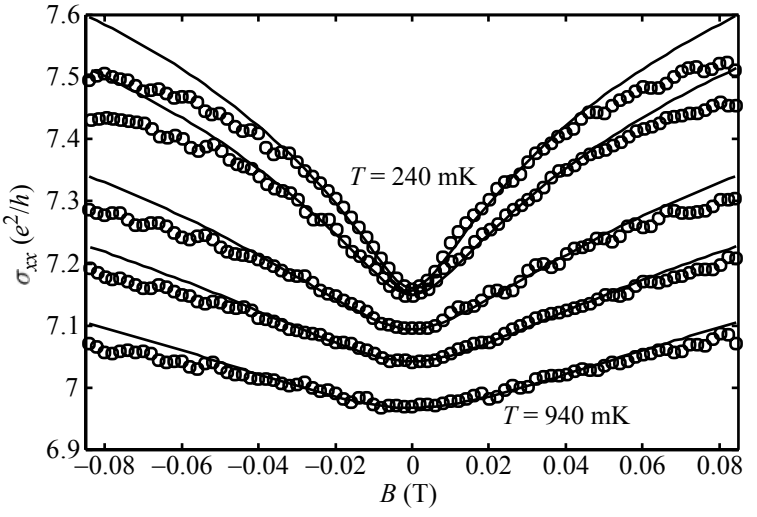


Fig. 15.6 Fit of the weak localization correction formula (15.6) to the measured data at different temperatures. From these fits, the phase coherence time τ_φ can be determined. (Reprinted with permission from Senz *et al.*, 2000*b*. Copyright 2000 by the American Physical Society.)

and the effect can therefore not be observed. The sample for which the data are shown in Fig. 15.5 has a mobility of about $7000\,\mathrm{cm}^2/\mathrm{Vs}$. The parameter $\sigma/(e^2/\hbar)$ is 1.2 and $f(1.2) = 0.53$. From the condition $\tau_\varphi \gg \tau_\mathrm{e}$ we obtain with eq. (15.3) $T \ll 3.3\,\mathrm{K}$ which can be easily achieved in the experiment. If the theoretical prediction (15.6) is fitted to the data, as shown in Fig. 15.6, the temperature-dependent phase coherence time $\tau_\varphi(T)$ can be extracted from the experiment. Figure 15.7 shows the linear temperature dependence which is expected according to eq. (15.3). The saturation at the lowest temperatures is probably due to the fact that the electron temperature is slightly higher than the lattice temperature. The prefactor $f(a)$ from eq. (15.3) is in this experiment about a factor of six larger than predicted theoretically. This deviation is relatively large. Typically deviations of about a factor 2 are found.

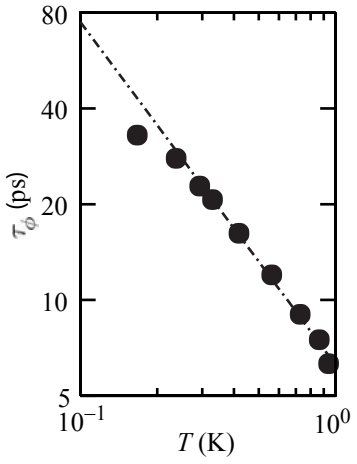


Fig. 15.7 Temperature-dependent phase coherence time as determined from the weak localization measurements. (Reprinted with permission from Senz *et al.*, 2000*a*. Copyright 2000 by the American Physical Society.)

Weak localization in three- and one-dimensional systems. The weak localization effect can not only occur in two-dimensional systems, but also in one or three dimensions. In three-dimensional systems the diffusive return probability is smaller than in two dimensions and the relative effect is weaker, but it can be observed, for example, in diffusive metallic conductors. In one-dimensional systems, where $l_\varphi > W$ the effect is much stronger than in two dimensions. The magnetic field dependence of the magnetoconductance has been calculated to be (Beenakker and van Houten, 1988*b*)

$$\delta G(B) = -\frac{2e^2}{h}\frac{1}{L}\left(\frac{1}{D\tau_\varphi} + \frac{1}{D\tau_B}\right)^{-1/2},$$

where $\tau_B = 3\hbar^2/(e^2W^2DB^2)$, D is the diffusion constant, and W and L are the width and the length of the sample, respectively.

15.5 Validity range of the Drude–Boltzmann theory

In chapter 10 we discussed the semiclassical description of electron transport in the framework of the Drude–Boltzmann model without caring about the range of validity of this description. We found the expression

$$\sigma_{xx} = \frac{ne^2\tau_\mathrm{e}}{m^\star} = \frac{2e^2}{h}\frac{E_\mathrm{F}}{\hbar/\tau_\mathrm{e}} = \frac{e^2}{h}k_\mathrm{F}l_\mathrm{e}$$

for the conductivity at zero magnetic field and low temperature, where τ_e is the elastic scattering time and l_e is the elastic mean free path. We calculated the elastic scattering time quantum mechanically from first order perturbation theory (Fermi's golden rule). This approximation will be appropriate if

$$\hbar/\tau_\mathrm{e} \ll E_\mathrm{F}.$$

This energy criterion is equivalent to the *Ioffe–Regel criterion*,

$$k_\mathrm{F}l_\mathrm{e} \gg 1.$$

It means that the mean free path l_e of the electrons is large compared to the Fermi wavelength $\lambda_\mathrm{F} = 2\pi/k_\mathrm{F}$. If l_e becomes comparable with λ_F, the wave functions tend to localize. In the case

$$k_\mathrm{F}l_\mathrm{e} < 1$$

we talk about *strong localization*. The influence of increasing potential fluctuations (or scattering) on the quantum states has been studied using a variety of models. Indeed, our previous discussion (section 15.1) of the weak localization effect reflects the influence of coherent scattering on the conductivity. The important result of all these theories is that potential fluctuations tend to localize the wave functions as shown

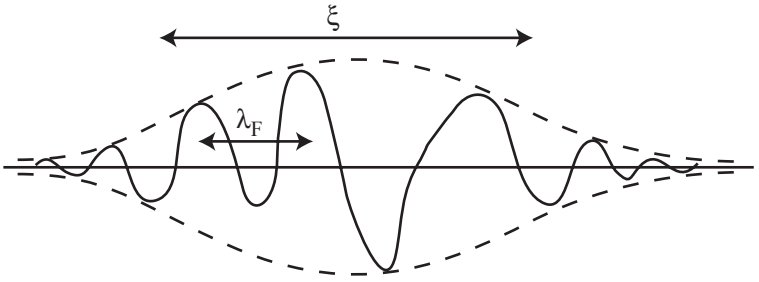


Fig. 15.8 Schematic representation of a wave function that is localized by spatial fluctuations of a potential. The wave function oscillates on the scale of the Fermi wavelength λ_F, but is localized by an envelope which decays exponentially from the region of localization and has a characteristic extent ξ.

schematically in Fig. 15.8. This means that there is a new length scale ξ called the *localization length*. Strong potential fluctuations lead to a small localization length. In the extreme case ξ becomes comparable to the Fermi wave length λ_F. Electrons are then localized around a single impurity site. At low temperatures, hopping transport will occur where electrons hop from one site to another either by thermal activation or by quantum tunneling. Weak potential fluctuations lead to a large localization length $\xi \gg \lambda_F$. Recently the question has arisen whether there might be quantum phase transition in two-dimensional electronic systems as a function of the strength of the fluctuating potential, such that below a critical strength all states are extended, but above it all states are localized. In the following we will discuss the basic physics background of this question without entering the current discussion of the topic. We will find that this discussion gives us a view on the weak localization effect, complementary to the microscopic picture elaborated earlier. Historically, the scaling approach to be discussed below triggered all the theoretical developments that eventually led to the microscopic picture that we discussed first for pedagogical reasons.

15.6 Thouless energy

The expression of the Drude conductivity can be written in the form of the zero-temperature Einstein relation (10.54)

$$\sigma_{xx} = \frac{ne^2\tau_e}{m^\star} = e^2\mathcal{D}_d(E_F)D_d,$$

if we introduce the diffusion constant $D_d = v_F^2\tau_e/d$ and the density of states at the Fermi energy E_F, $\mathcal{D}_d(E_F)$ in d-dimensions ($d = 1, 2, 3$). We now ask for the conductance $G(L)$ of a cubic ($d = 3$, square $d = 2$, stretch $d = 1$) piece of material with side length L which is much bigger than the mean free path l_e of the electrons. We obtain

$$G(L) = \sigma_{xx}\frac{L^{d-1}}{L} = \frac{e^2}{\hbar}(\mathcal{D}_d(E_F)L^d)\frac{\hbar D_d}{L^2}. \qquad (15.7)$$

In this expression, the quantity $\mathcal{D}_d(E_F)L^d := 1/\Delta$ is the number of quantum states at the Fermi energy per unit energy interval, and we can introduce a typical energy level spacing Δ. The quantity $\hbar D_d/L^2$

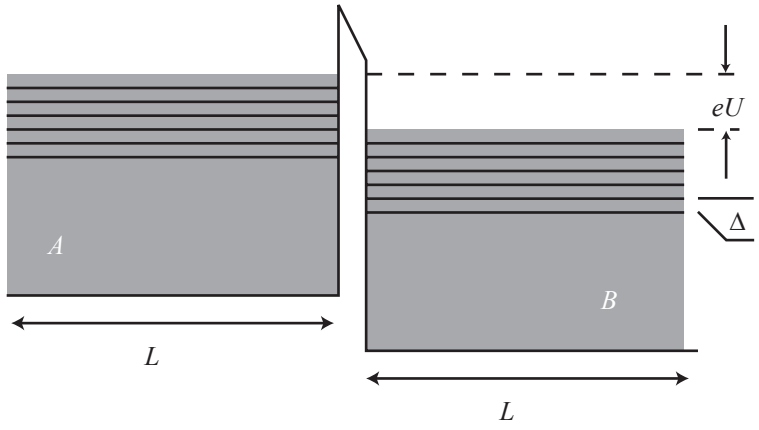


Fig. 15.9 Schematic representation of two blocks A and B of material with extent L, coupled through a thin (rough) tunneling barrier. In each of the two blocks of material the mean level spacing is given by Δ.

has the dimension of energy. It is called the *Thouless energy* (Thouless, 1977)

$$E_{\mathrm{Th}} = \frac{\hbar D_d}{L^2}. \tag{15.8}$$

The physical meaning of the Thouless energy becomes clear if we realize that $D_d/L^2 = 1/\tau_{\mathrm{Th}}$ is the inverse of a time scale τ_{Th} telling us how long an electron needs in order to explore the area L^2 diffusively, or in other words, at what average rate the diffusing electron reaches the boundary of the square. The Thouless energy corresponds to this time scale via Heisenberg's uncertainty relation. In the picture of wave functions it tells us how sensitive the energy of a particular wave function is to a change in the boundary conditions at the square boundary. It is intuitively clear that a strongly localized state will hardly change when the sample conditions at the sample edge are changed (E_{Th} small), whereas an extended state will change strongly (E_{Th} large).

The conductance of the square of material can therefore be written as

$$G(L) = \frac{e^2}{\hbar}\frac{E_{\mathrm{Th}}(L)}{\Delta(L)}, \tag{15.9}$$

i.e. the conductance is equal to the conductance quantum times the number of energy levels within the energy interval E_{Th}. The dimensionless conductance $g(L) = G(L)/(2e^2/h)$ is frequently called the *Thouless number*. In the range of validity of the Drude–Boltzmann theory we have $g(L) \gg 1$, i.e., $\Delta(L) \ll E_{\mathrm{Th}}(L)$. In the regime of strong localization $g(L) \ll 1$ and, correspondingly, $E_{\mathrm{Th}} \ll \Delta$. We can estimate the localization length via the condition $g(\xi) \approx 1$, if an expression for $g(L)$ is known.

Another approach to clarify the meaning of the Thouless energy (Imry, 2002) considers two square pieces of material A and B that are separated by a thin tunneling barrier as shown schematically in Fig. 15.9. We assume that the barrier is rough, such that the momentum along the barrier is not conserved for tunneling processes and consequently every state in A couples to every other state in B with similar strength. We

express this coupling by an average squared matrix element $\overline{t^2}$. The tunneling rate for an electron from a certain state in A into any state in B is then given by Fermi's golden rule as

$$\frac{1}{\tau} = \frac{2\pi}{\hbar}\overline{t^2}\mathcal{D}_d(E_\mathrm{F}).$$

If we apply a small voltage U across the tunneling barrier, $e\mathcal{D}_d(E_\mathrm{F})L^dU$ states contribute to the tunneling current, which decay with the typical time constant τ from A to B. The tunneling current is therefore $I = e^2\mathcal{D}_dL^dU/\tau$ and the conductance is

$$G(L) = \frac{e^2\mathcal{D}_d(E_\mathrm{F})L^d}{\tau} = \frac{e^2}{\hbar}(\mathcal{D}_d(E_\mathrm{F})L^d)\frac{\hbar}{\tau}.$$

Comparing with eqs (15.7) and (15.8) leads us to identify

$$E_\mathrm{Th} = \frac{\hbar}{\tau} = 2\pi\overline{t^2}\mathcal{D}_d(E_\mathrm{F}) = 2\pi\frac{\overline{t^2}/L^d}{\Delta}.$$

In the limiting case of weak tunneling coupling, $\overline{t^2}/L^d \ll \Delta$, the Thouless energy is simply the lifetime broadening of an energy level. In this case every level stays essentially localized in its own block. If the coupling strength is stronger, and even stronger than Δ, the localization of states disappears and the states become more and more extended over both blocks of material. In this limit one can see E_Th as the characteristic energy scale over which states of the two material blocks admix to form a particular state of the coupled system. We can then say that states within the interval E_Th are correlated. Sometimes the Thouless energy is therefore called the *correlation energy*.

We briefly summarize the results of the above discussion: The ratio E_Th/Δ of two energy scales, i.e., the sensitivity of changes in the boundary condition divided by the mean level spacing, is a dimensionless measure of the coupling of two quantum systems and is equal to the dimensionless conductivity $g(L)$. For $g(L) \gg 1$ neighboring blocks of material are strongly coupled and the states are extended over both. For $g(L) \ll 1$ the two blocks are essentially decoupled and the states are localized in either one or the other of the two. The size of the localization length ξ can therefore be estimated from $g(\xi) \approx 1$.

15.7 Scaling theory of localization

The conductance of a macroscopic disordered sample can usually be calculated neither analytically nor numerically. The goal of the *scaling theory* of localization is the calculation of the conductance of a macroscopic sample. The basic idea is the following: the sample is considered to be divided in smaller blocks of material with side length $L \gg l$ for which, for example, a numerical calculation of $G(L)$ is possible (Fig. 15.10). The conductance of the macroscopic sample is then deduced from that of the small cube by subsequent doubling of the cube's side length. This

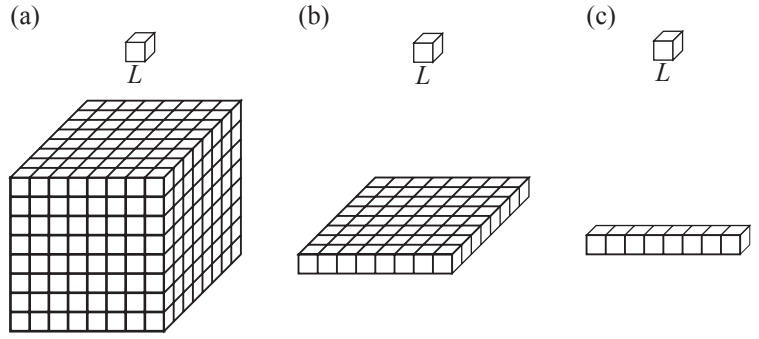


Fig. 15.10 Basic idea of the scaling theory. A macroscopic sample is considered to be divided into smaller cubes for which the conductance can be calculated. Using the scaling function, the conductance of the macroscopic sample can be determined from the conductance of the small cube. (a) three-dimensional case, (b) two-dimensional case, (c) one-dimensional case.

procedure can be applied in one-, two-, or three-dimensions. The important ingredient for this to work is that the conductance $G(2L)$ of a piece of material with side length $2L$ depends only on the conductance $G(L)$. Mathematically this is expressed by the relation

$$\frac{d\ln g}{d\ln L} = \beta(g)$$

for the dimensionless conductivity g, where $\beta(g)$ is called the *scaling function*, and $g = E_{\mathrm{Th}}/\Delta$ is called the *scaling parameter*.

How can we see that it is reasonable to use g as the only scaling parameter? Assuming that we have solved Schrödinger's eigenvalue problem for two material blocks of size L such that the energy levels and wave functions are known, we find the solutions for the combined system by matching the wave functions at the boundaries where the two blocks touch. States of the two blocks will mix to form extended states if there are many levels in an energy interval E_{Th} in both blocks, i.e., if $E_{\mathrm{Th}}/\Delta \gg 1$. In turn, states of the two blocks will hardly mix if there are few states within E_{Th}, i.e., if $E_{\mathrm{Th}}/\Delta \ll 1$. Therefore it is reasonable to assume that $g(2L)$ is essentially given by $E_{\mathrm{Th}}/\Delta = g(L)$.

The question is now, how the scaling function $\beta(g)$ can be determined. In order to get a feeling for this function, we consider a few well-known limiting cases. If $g \gg 1$ we are in the limit of the Drude–Boltzmann theory and have

$$g(L) = \frac{h}{2e^2}\sigma L^{d-2}, \tag{15.10}$$

where $d = 1, 2, 3$ is the dimensionality of the conductor, and σ is the specific conductivity. In this limit we find

$$\lim_{g\to\infty} \beta(g) = d - 2,$$

i.e., β tends to a constant in all dimensions.

On the other hand, it is known that the conductance depends exponentially on the sample size in the regime of strong localization, i.e.,

$$g(L) = g_0 e^{-L/\xi}.$$

This leads to

$$\lim_{g\to 0} \beta(g) = -\frac{L}{\xi} = \ln\frac{g}{g_0}$$

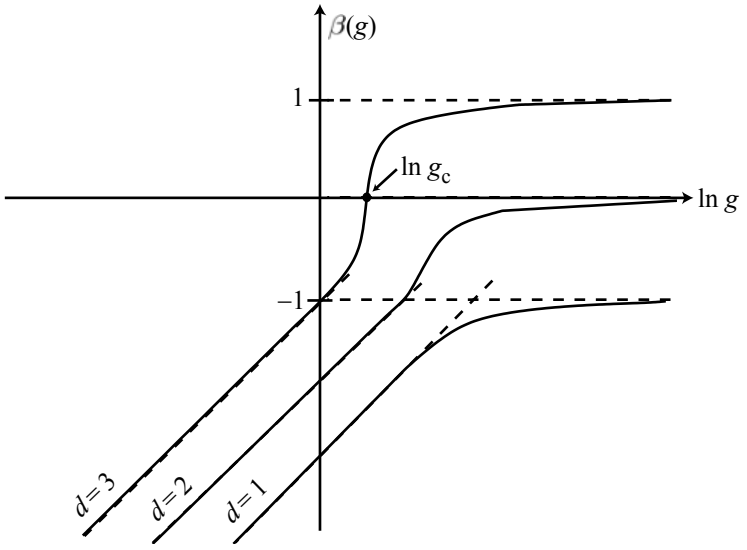


Fig. 15.11 Scaling function $\beta(g)$ and its limiting cases plotted versus $\ln g$ for one-, two-, and three-dimensional systems.

in all dimensions.

From these two examples we can see that the scaling function β depends in both limiting cases only on g. Scaling theory assumes that between these two limiting cases, the scaling function β does also depend only on the one scaling parameter g. In Fig. 15.11 the scaling function is schematically shown as a function of $\ln g$ as was suggested in Abrahams *et al.*, 1979.

We now discuss qualitatively what the suggested functional form of $\beta(g)$ means for systems of different dimensions. Suppose we have calculated $g(L)$ for a small system of size L. We then obtain the dimensionless conductance $g(L_\mathrm{m})$ of the macroscopic system of size L_m from solving the integral equation

$$\int_{\ln g(L)}^{\ln g(L_\mathrm{m})} \frac{d\ln g}{\beta(g)} = \int_{\ln L}^{\ln L_\mathrm{m}} d\ln L = \ln\frac{L_\mathrm{m}}{L}. \qquad (15.11)$$

For a one-dimensional system, $\beta(g) < 0$ for all values of g. This means that $g(L_\mathrm{m}) < g(L)$, no matter from which value $g(L)$ we start. The scaling theory therefore predicts that macroscopic one-dimensional systems are always insulators, because for arbitrary L_m, $g(L_\mathrm{m})$ can become arbitrarily small.

The situation is different in the case of a three-dimensional system. For certain values of $g(L) > g_\mathrm{c}$, β is a positive number, whereas for $g(L) < g_\mathrm{c}$ it is negative (Fig. 15.11). This means that the conductance of a macroscopic three-dimensional system can become very large with increasing system size, if $g(L)$ is bigger than g_c. We call this *metallic behavior*. On the other hand, the conductance can become arbitrarily small, if $g(L) < g_\mathrm{c}$, and we call it an *insulator*. In three-dimensional systems there is a *metal–insulator transition* as a function of the strength of potential fluctuations in the sample.

The case of the two-dimensional system is very interesting because the limit of β is zero for $g \to \infty$. It is therefore very important to find out whether β is positive or negative for large g. In the first case, we would expect a metal–insulator transition even in two-dimensional systems, whereas in the second case, all macroscopic two-dimensional systems would be insulating. Abrahams *et al.*, 1979, suggested a result that was later proven to be correct. According to this theory, the scaling function in two dimensions has, for large g, the asymptotic form

$$\beta(g) = -\frac{g_0}{g},$$

where $g_0 \approx 1$. The prediction is therefore that in two dimensions, $\beta(g) < 0$, i.e., all macroscopic systems are insulators and there is no metal–insulator transition in two dimensions. Using eq. (15.11) we find

$$g(L_\mathrm{m}) = g(L) - \ln \frac{L_\mathrm{m}}{L}.$$

The logarithmic term makes the result different from the Drude–Boltzmann limit (15.10) for $d = 2$. In fact this term reminds us of the logarithmic correction of the Drude conductivity found as a result of enhanced coherent backscattering [cf., eq. (15.1)]. If we insert eq. (15.10) for $g(L)$ we obtain

$$G(L_\mathrm{m}) = \sigma - \frac{2e^2}{h} \ln \frac{L_\mathrm{m}}{L}. \tag{15.12}$$

In two-dimensional systems there is a logarithmic correction to the Drude–Boltzmann conductivity with its origin in the quantum diffusion of electrons. The above equation can, however, only be used as long as the logarithmic term is small compared to the Drude–Boltzmann σ.

From the above considerations we can estimate the localization length ξ for a two-dimensional system using the relation $g(\xi) \approx 1$. We obtain

$$\xi \approx l_\mathrm{e} e^{k_\mathrm{F} l_\mathrm{e}/2}. \tag{15.13}$$

Here we have assumed that $k_\mathrm{F} l_\mathrm{e} \gg 1$, and we have set $L \approx l_\mathrm{e}$ which gives only a small error, because of the logarithm. We see that the localization length grows exponentially with $k_\mathrm{F} l_\mathrm{e}$. If we consider a high-mobility electron gas with a density $n = 3 \times 10^{15}\,\mathrm{m}^{-2}$ and a low-temperature mobility $\mu = 10^6\,\mathrm{cm}^2/\mathrm{Vs}$, we find $k_\mathrm{F} l_\mathrm{e} = 1241$ and $l_\mathrm{e} = 9\,\mu\mathrm{m}$. This leads to $\xi \approx 2 \times 10^{264}\,\mathrm{m}$ which is an astronomically large length scale. In turn, for an electron gas with the same density, but a mobility $\mu = 10^4\,\mathrm{cm}^2/\mathrm{Vs}$, we find $k_\mathrm{F} l_\mathrm{e} = 12.4$, $l_\mathrm{e} = 90\,\mathrm{nm}$ and $\xi \approx 44\,\mu\mathrm{m}$. For even smaller mobilities, the regime of strong localization is quickly reached.

What is now the significance of the phase coherence length for the localization of electrons in two-dimensional electron gases? Obviously the quantum mechanical conductance can only be scaled up to the coherence length l_φ. Above this length scale the quantum description breaks down and the conductances are scaled according to classical laws. However, classically the conductance G in two dimensions does not change when the system size is doubled. The reason is that the conductances of two

blocks of material attached in parallel will add doubling the conductance, for the subsequent connection in series, however, the resistances have to be added and the conductance is again reduced by the same factor of two. Using eq. (15.12) we therefore find for the quantum mechanical conductance σ_{qm} (in contrast to the Drude conductance σ)

$$\frac{\delta\sigma_{\mathrm{qm}}(T)}{\sigma} = -\frac{1}{k_{\mathrm{F}} l_{\mathrm{e}}} \ln \frac{l_\varphi(T)}{l_{\mathrm{e}}}. \tag{15.14}$$

The coherent quantum diffusion of electrons leads to a logarithmic correction of the classical Drude conductivity. This result reproduces our earlier finding expressed in eq. (15.1).

15.8 Length scales and their significance

In the Drude–Boltzmann theory the Fermi wavelength λ_{F}, and the elastic mean free path l_{e} are of importance. Furthermore we have now introduced the localization length ξ and the phase-coherence length l_φ. In general, if we study electronic transport phenomena, we have to compare these length scales with the (lateral) system size L. We can classify transport phenomena by relating all these length scales.

The weak localization regime. If the elastic mean free path l_{e} in a two-dimensional system is large compared to the Fermi wavelength λ_{F}, according to eq. (15.13) $\xi \gg l_{\mathrm{e}}$. If furthermore $l_\varphi \gg l_{\mathrm{e}}$, but $l_\varphi \ll \xi$, then the logarithmic correction of the Drude conductance is important and eq. (15.14) is relevant. This is called the *weak localization regime*. The hierarchy of length scales is given by $L, \xi \gg l_\varphi \gg l_{\mathrm{e}} > \lambda_{\mathrm{F}}$. The electron motion is governed by quantum diffusion. This case arises only at low temperatures, because there, $l_\varphi(T)$ is sufficiently large. If $l_\varphi, l_{\mathrm{e}} \gg \lambda_{\mathrm{F}}$, the electron motion is called semiclassical meaning that the electrons follow essentially classical trajectories, but carry quantum phase information which makes them susceptible to interference. Also the case $\lambda_{\mathrm{F}} \approx l_{\mathrm{e}}$ is in the range of weak localization, but here the electron motion is no longer semiclassical, but rather follows the rules of true quantum diffusion.

Diffusive classical transport. If $L, \xi \gg l_{\mathrm{e}} \geq l_\varphi \gg \lambda_{\mathrm{F}}$, the logarithmic quantum correction to the conductivity does not play a role, the electron motion is classically diffusive, and the conductivity is well described by the Drude model. This scenario applies, for example, at elevated temperatures where the phase coherence length $l_\varphi(T)$ is small. This case also occurs in extremely pure two-dimensional electron gases having mean free paths at the lowest temperatures that are comparable or larger than l_φ.

Quantum regime of strong localization. The relations $\xi \approx l_{\mathrm{e}} \approx \lambda_{\mathrm{F}}$ describe the regime of strong localization of an electron gas. At low temperatures $l_\varphi > \xi, l_{\mathrm{e}}, \lambda_{\mathrm{F}}$ and transport is coherent. The diffusive Drude

model is not appropriate here, but hopping transport occurs between localization sites. Also, in this regime of electron transport, interference corrections of the conductivity can arise, originating from the interference of alternative hopping paths.

The regime of classical strong localization. Also, in the regime of the strong localization, quantum interference becomes irrelevant if $l_\varphi \leq \xi$. We can then talk about incoherent hopping transport, classical localization, and classical percolation.

Mesoscopic systems. The regimes of electronic transport identified above are valid for two-dimensional electron gases where the system size L is much larger than any other length scale. Now we consider systems in which the system size is comparable or even smaller than relevant length scales of electron transport. Such systems are called *mesoscopic systems*. The dimensionality of a mesoscopic system is defined by comparing the phase coherence length l_φ with the system size. For example, if the width W of a Hall bar is smaller than l_φ, but its length L larger, then we talk about a one-dimensional mesoscopic system. For $L, W > l_\varphi$ the mesoscopic system is called two-dimensional; if $L, W < l_\varphi$ it is called zero-dimensional.

Diffusive mesoscopic systems. If in a one- or zero-dimensional mesoscopic system, the system size $L, W \gg l_\mathrm{e}$, we call electron transport in this mesoscopic system diffusive.

Ballistic mesoscopic systems. If in turn in a one- or zero-dimensional system $W, L \ll l_\mathrm{e}$, then we talk about ballistic electron transport. Scattering of electrons at sample boundaries dominates in this case over scattering at spatial potential fluctuations.

Quasi-ballistic systems. The regime between diffusive and ballistic mesoscopic systems is sometimes called quasi-ballistic. In this regime, for example, $W < l_\mathrm{e}$, but $L > l_\mathrm{e}$.

15.9 Weak antilocalization and spin–orbit interaction

In the presence of strong spin–orbit interaction (SOI), as it is found, for example, in n-InAs, n-InSb, or in p-GaAs the phase-coherent backscattering is altered in a very interesting way. As the partial waves travel along diffusive closed loops, the spin is rotated under the influence of the SOI. It was shown in Hikami *et al.*, 1980, that SOI can reverse the sign of the weak localization correction to the conductivity compared to the case in which SOI is absent, and that the magnitude of the correction is reduced by a factor of 1/2. In order to understand this effect at

least qualitatively, we follow the reasoning of Bergmann 1982. Essentially we have to generalize the considerations made for phase-coherent backscattering without spin by including the spin degree of freedom. Spin–orbit interaction can lead to spin rotation during scattering and between scattering events. A way to visualize this effect is to regard the spin as diffusing on the Bloch sphere as shown in Fig. 15.12. If the partial wave starts the path in a particular spin state $|s\rangle$, it will arrive after one clockwise revolution around the loop (see Fig. 15.1) in a state

$$|s'\rangle = R\,|s\rangle\ ,$$

where the rotation operator R is the product of a large number of small rotations occurring subsequently, i.e.,

$$R = R_n \ldots R_2 R_1.$$

If we consider the partial wave propagating along the time-reversed path, the sequence of the rotations is reversed, but the rotation angle of each individual section of the path is also inverted. Therefore, after one counterclockwise revolution around the same loop it will arrive in the state

$$|s''\rangle = \tilde{R}\,|s\rangle\ ,$$

where

$$\tilde{R} = \tilde{R}_1 \tilde{R}_2 \ldots \tilde{R}_n,$$

with $R_i \tilde{R}_i = 1$ meaning that $\tilde{R}_i = R_i^{-1}$. In general, the rotation operator for a spin can be written in matrix form as

$$R(\alpha,\beta,\gamma) = \begin{pmatrix} \cos\frac{\alpha}{2} e^{i(\beta+\gamma)/2} & i\sin\frac{\alpha}{2} e^{-i(\beta-\gamma)/2} \\ i\sin\frac{\alpha}{2} e^{i(\beta-\gamma)/2} & \cos\frac{\alpha}{2} e^{-i(\beta+\gamma)/2} \end{pmatrix},$$

where α, β, γ are the Euler angles. This matrix has the property that the inverse rotation

$$R_i^{-1}(\alpha_i,\beta_i,\gamma_i) = \tilde{R}_i(\alpha_i,\beta_i,\gamma_i) = R_i^{\mathrm{T}}(-\alpha_i,-\beta_i,-\gamma_i) = R_i^{\dagger}(\alpha_i,\beta_i,\gamma_i),$$

meaning that it is unitary. It is straightforward to show that R (and $\tilde{R}$) are unitary if all the R_i are unitary.

Interference of the two time-reversed paths requires us to calculate

$$(\langle s'| + \langle s''|\,)\,(|\,s'\rangle + |\,s''\rangle\,) = 2 + \langle s'|s''\rangle + \langle s''|s'\rangle\,.$$

For the interference contribution, we find

$$\langle s''|s'\rangle = \left\langle \tilde{R}s \,|\, Rs \right\rangle = \left\langle R^{\dagger}s \,|\, Rs\right\rangle = \left\langle s\,\middle|R^2\middle|\,s\right\rangle.$$

We can introduce a characteristic time scale τ_{SO} describing the randomization of the spin direction in time due to SOI. For times $t \ll \tau_{\mathrm{SO}}$ the spin orientation has not changed significantly, whereas for times $t \gg \tau_{\mathrm{SO}}$ the spin orientation is completely randomized.

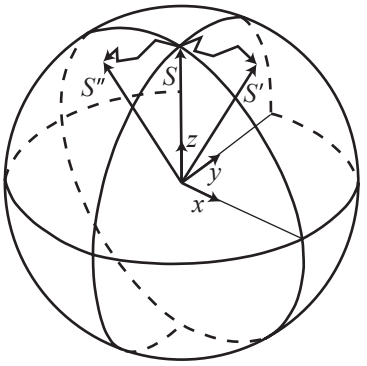


Fig. 15.12 Spin diffusion on the Bloch sphere. (Reprinted from Bergmann, 1982 with permission from Elsevier.)

In the case of weak spin–orbit interaction, $\tau_{\rm SO} \gg \tau_\varphi$, the spin will stay essentially polarized in the same direction throughout the evolution, R becomes the unity matrix in good approximation, because all angles are close to zero, and as a result quantum backscattering is *enhanced* by a factor of two compared to the classical backscattering probability. This reproduces the case of weak localization that we considered in the previous section.

More interesting is the case of strong spin–orbit interaction, $\tau_{\rm SO} \ll \tau_\varphi$, where the spin polarization gets completely randomized as the electron travels along a typical path. However, on the time-reversed path the spin experiences exactly the opposite rotation. In this case we have to calculate the expectation value of R^2. The square of the rotation matrix is given by

$$R^2(\alpha,\beta,\gamma) = \begin{pmatrix} \cos^2\frac{\alpha}{2}e^{i(\beta+\gamma)} - \sin^2\frac{\alpha}{2} & \frac{i}{2}\sin\alpha e^{-i\beta}\left(1+e^{i(\beta+\gamma)}\right) \\ \frac{i}{2}\sin\alpha e^{-i\gamma}\left(1+e^{i(\beta+\gamma)}\right) & \cos^2\frac{\alpha}{2}e^{-i(\beta+\gamma)} - \sin^2\frac{\alpha}{2} \end{pmatrix}.$$

If the spinor $|\,s\rangle = (a,b)$ we find

$$\langle s'|s''\rangle = \cos^2\frac{\alpha}{2}\left(e^{i(\beta+\gamma)}|a|^2 + e^{-i(\beta+\gamma)}|b|^2\right) - \sin^2\frac{\alpha}{2} + \frac{i}{2}\sin\alpha\left[ab^\star\left(e^{-i\beta}+e^{i\gamma}\right) + a^\star b\left(e^{i\beta}+e^{-i\gamma}\right)\right].$$

If we average this interference term over many pairs of time-reversed paths in a diffusive sample (this amounts to averaging over all possible angles α,β,γ), all terms in the above expression will average out, except the term $-\sin^2\alpha/2$ which gives an average contribution of $-1/2$. We therefore conclude that strong spin–orbit scattering leads to

$$(\langle s'| + \langle s''|)\,(|\,s'\rangle + |\,s''\rangle) = 2 - \frac{1}{2} - \frac{1}{2} = 1.$$

This implies that strong spin–orbit interaction *reduces* the quantum backscattering probability to one half of the classical backscattering probability, because destructive interference dominates.

Low-field magnetoresistance. The magnetoresistance in the presence of spin–orbit scattering can be described in the same spirit as before if we introduce the action of $\tau_{\rm SO}$ into the expression for the probability distribution (15.5) of areas. Introducing the additional factor (Chakravarty and Schmid, 1986)

$$\frac{1}{2}\left(3e^{-4S/3l_{\rm SO}^2} - 1\right),$$

with $l_{\rm SO} = D\tau_{\rm SO}$, into the function $P(S,l_\varphi)$ in eq. (15.5) has the desired effect: for areas $S \gg l_{\rm SO}^2$ the spin is completely randomized and the above factor is $-1/2$. In the opposite case, $S \ll l_{\rm SO}^2$, the spin is not rotated and the above factor is one. Therefore, depending on the area S this factor switches from the usual weak localization behavior

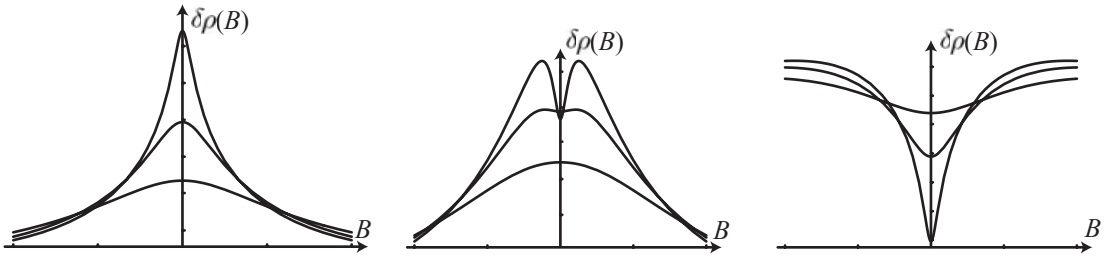


Fig. 15.13 Magnetoresistance correction calculated from eq. (15.4) for different strengths of the spin–orbit interaction. (a) $l_{\mathrm{SO}}/l_{\mathrm{e}} = 30$, (b) $l_{\mathrm{SO}}/l_{\mathrm{e}} = 3$, (c) $l_{\mathrm{SO}}/l_{\mathrm{e}} = 1$. In each subfigure, three curves are shown for the parameters $l_\varphi/l_{\mathrm{e}} = 5$, $l_\varphi/l_{\mathrm{e}} = 2.5$, and $l_\varphi/l_{\mathrm{e}} = 1.5$.

(small areas) to the weak antilocalization behavior. Therefore we use in eq. (15.4) the expression

$$P(S, l_\varphi)dS \propto \frac{1}{4\pi S}\frac{1}{2}\left(3e^{-4S/3l_{\mathrm{SO}}^2} - 1\right)\left(1 - e^{-S/l_{\mathrm{e}}^2}\right)e^{-S/l_\varphi^2}dS$$

for calculating the correction to the conductivity.

The result of such a calculation is shown in Fig. 15.13. It can be seen that a weak localization maximum in the magnetoresistance is recovered for $l_{\mathrm{SO}} > l_\varphi > l_{\mathrm{e}}$ in Fig. 15.13(a). In the opposite case of $l_{\mathrm{SO}} = l_{\mathrm{e}} < l_\varphi$ a weak antilocalization minimum is found [Fig. 15.13(c)]. In the intermediate case, where $l_{\mathrm{e}} < l_{\mathrm{SO}} < l_\varphi$, the weak localization peak develops an antilocalization dip at small fields [Fig. 15.13(b)]. In all cases, the phase coherence length l_φ determines the region of strongest curvature on a field scale $B_\varphi = \phi_0/l_\varphi^2$ around zero field, whereas the elastic mean free path l_{e} determines the width of the maximum by the magnetic field scale $B_{\mathrm{e}} = \phi_0/l_{\mathrm{e}}^2$. The spin–orbit length l_{SO} introduces an additional field scale $B_{\mathrm{SO}} = \phi_0/l_{\mathrm{SO}}^2$ which marks the turnover from weak antilocalization to weak localization in Fig. 15.13(b).

Spin–scattering mechanisms. So far we have introduced the spin–orbit scattering time τ_{SO} and the corresponding spin–orbit scattering length l_{SO} heuristically without discussing the underlying physical spin-relaxation mechanisms. In the following we will give a brief discussion of the most relevant mechanism. The so-called *D'yakonov–Perel mechanism* (D'yakonov and Perel, 1971), also called skew scattering, is believed to be dominant in most III-V semiconductors lacking inversion symmetry and low-dimensional structures fabricated from these materials (with the exception of narrow band gap materials with large separation of the spin–orbit split-off band from the valence band maximum, like InSb). It is caused by the following scenario: The effective magnetic field (9.16) that the spin of an electron experiences depends on the wave vector $\mathbf{k}$ of the orbital state. If an electron has scattered at an impurity from one $\mathbf{k}$-state at the Fermi surface to another, it will be subject to a spin–orbit-induced magnetic field with different orientation and magnitude. This means that after each scattering event the spin rotates around a different

precession axis with a different angular velocity. Within the typical time span τ_e between two elastic scattering events, the spin will precess by the typical angle $\delta\varphi = \bar{\omega}_L \tau_e$, where $\bar{\omega}_L$ is the typical Larmor frequency. In typical materials the precession angle $\delta\varphi(\tau_e) \ll 1$ for electrons in the conduction band. After a large number of random scattering events, the spin orientation will have performed a random walk on the Bloch sphere and the correlation between the initial spin orientation and the final spin orientation is lost. To be more specific, after a time $t \gg \tau_e$, a number t/τ_e steps of the random walk have occurred and the orientation will have acquired an average spread $\langle \delta\varphi^2(t) \rangle = \delta\varphi^2(\tau_e) t/\tau_e = \bar{\omega}_L^2 \tau_e t$. We now estimate the spin–orbit scattering rate as the average time it takes for the spread of the phase to become of the order 1. This leads to

$$\frac{1}{\tau_{SO}} = \bar{\omega}_L^2 \tau_e.$$

The spin–orbit relaxation rate is proportional to the scattering time. This implies that increasing disorder (smaller τ_e) leads to *larger* spin–orbit time τ_{SO}, which is a bit counterintuitive. On the other hand, stronger spin–orbit interaction would lead to a larger value of the typical $\bar{\omega}_L$ and therefore to a larger spin-relaxation rate, as expected. The D'yakonov–Perel mechanism assumes that the scattering event itself does not alter the spin orientation.

Other spin-relaxation mechanisms are the *Bir–Aronov–Pikus* mechanism, the *hyperfine interaction*, and the *Elliott–Yafet mechanism* (Elliott, 1954). The Bir–Aronov–Pikus mechanism is important in p-doped semiconductors. The electron–hole exchange interaction causes fluctuating local magnetic fields that act on spin states. The hyperfine interaction is relevant in semiconductors in which the constituents of the nuclei possess a nonzero magnetic moment. Spin-flip scattering of electrons at nuclei can randomize the spin. The Elliott–Yafet mechanism assumes that the spin polarization is unaltered between elastic scattering events, but spin-rotation occurs during impurity scattering events, complementary to the Dyakonov–Perel mechanism. An in-depth discussion of all these spin relaxation mechanisms is beyond the scope of this book, but an excellent discussion can be found, for example, in Zutic *et al.*, 2004.

Experimental observation of the weak antilocalization effect. For the observation of the weak antilocalization effect, diffusive samples with sufficiently strong spin–orbit interaction are required. For example, electron gases in InAs or InGaAs are well suited for such experiments. Electron mobilities well below $10\,\mathrm{m}^2/\mathrm{Vs}$ help to see an appreciable effect. Figure 15.14 shows a measurement performed on an $Al_xGa_{1-x}As/In_xGa_{1-x}As/GaAs$ 13 nm pseudomorphic quantum well. The weak antilocalization enhancement of the zero field conductivity is seen to coexist with a weak localization reduction of the conductivity which becomes dominant at $|B| > 5\,\mathrm{mT}$. Theoretical curves fit the experimentally acquired data quite well in the range of applicability (between the two dotted vertical lines).

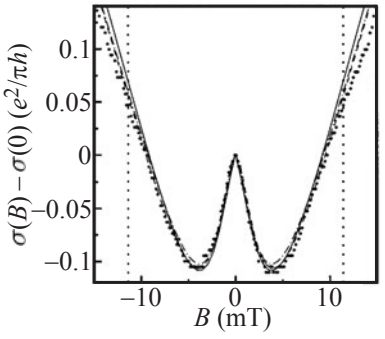


Fig. 15.14 Measurement of the weak antilocalization effect in an InGaAs quantum well with a two-dimensional electron gas (density $1.34 \times 10^{12}\,\mathrm{cm}^{-2}$, $\mu = 1.94\,\mathrm{m}^2/\mathrm{Vs}$). Circles indicate the measured points, all lines are theoretical fits. (Reprinted with permission from Knap *et al.*, 1996. Copyright 1996 by the American Physical Society.)

Fig. 15.15 shows a set of measurements performed on four different $In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.47}As$ quantum well samples where the strength of the spin–orbit interaction was increased from bottom to top by changing the inversion asymmetry of the quantum well through varying the remote doping profile. The sample that led to the lowest curve did not exhibit a weak antilocalization minimum at zero magnetic field, but rather a weak localization maximum, and increasing spin–orbit interaction strength leads to the development of a zero-field minimum which dominates over the weak localization maximum completely in the topmost trace.

The spin–orbit interaction is also very important in the valence band at the Γ-point, where it already leads to a splitting of the six-fold degenerate dispersion of hole states into a four-fold degenerate heavy hole/light hole band, and the spin–orbit split-off band in the bulk band structure of III-V semiconductors. Confinement to two-dimensions splits the degeneracy of heavy and light hole states such that, in low-density hole gases, only the two-fold degenerate heavy hole states are occupied. The in-plane dispersion of these heavy hole states is again split by the presence of spin–orbit interaction leading to two spin-split heavy hole dispersion branches (Winkler, 2003). As a consequence, a weak antilocalization effect does also occur, for example, in *p*-type GaAs two-dimensional hole gases.

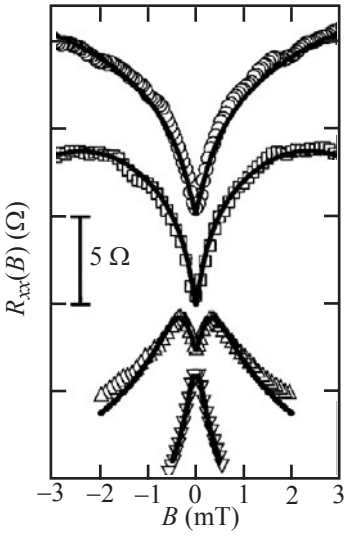


Fig. 15.15 Crossover from weak localization (bottom) to weak antilocalization (top) in $In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.47}As$ quantum well samples with differing spin–orbit interaction strength. Typical mobilities are $5\,\mathrm{m}^2/\mathrm{Vs}$ at densities $1 \times 10^{12}\,\mathrm{cm}^{-2}$. (Nitta and Koga, 2003. With kind permission from Springer Science and Business Media.)

Further reading

- Weak localization: Beenakker and van Houten 1991.
- Spin–orbit interaction and weak antilocalization: Winkler 2003.
- Papers: Bergmann 1984; Chakravarty and Schmid 1986; Knap *et al.* 1996; Zutic *et al.* 2004; Fabian and Sarma 1999.

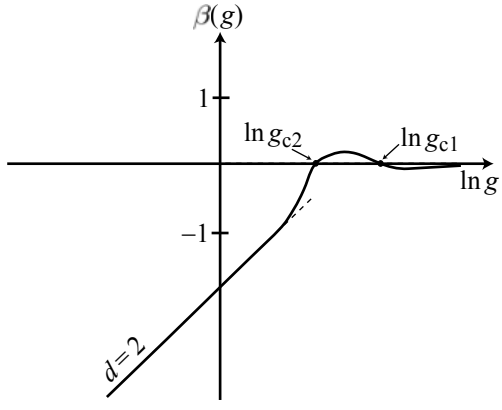
Exercises

(15.1) The lowest temperatures reached in transport experiments are of the order of 10 mK. Consider two Ga[Al]As heterostructures with identical electron densities $n_s = 3 \times 10^{15}\,\mathrm{m}^{-2}$, but different mobilities of $\mu_1 = 3\,\mathrm{m}^2/\mathrm{Vs}$ and $\mu_2 = 100\,\mathrm{m}^2/\mathrm{Vs}$ cooled to this temperature.

(a) Calculate the mean free paths l_1 and l_2 in the two systems. Estimate the localization lengths ξ_1 and ξ_2, and the phase-coherence lengths $l_{\varphi 1}$ and $l_{\varphi 2}$. Can the weak localization phenomenon be observed in both samples?

(b) Estimate how big a mesoscopic structure fabricated from one of these wafers can be in order to to have ballistic transport. Is it possible to realize a diffusive mesoscopic system with the high mobility sample?

(15.2) Assume that the scaling function $\beta(g)$ in two dimensions has the form depicted. Discuss the existence of a quantum phase transition in this case. What is the physical meaning of the two dimensionless conductances g_{c1} and g_{c2}?



Magnetotransport in two-dimensional systems

16

16.1 Shubnikov–de Haas effect

In chapter 10 the quantization of states under the influence of the external magnetic field was neglected. Such an approach is appropriate for small magnetic fields for which $\omega_c\tau \ll 1$. With increasing magnetic field the quantization of states leads to an oscillatory magnetoresistance that is seen if either the electron density or the magnetic field strength is changed. In two-dimensional electron gases this effect was measured for the first time by Fowler *et al.*, 1966. They investigated silicon MOS structures in a Corbino geometry. The resulting conductance, measured at a constant magnetic field of 3.3 T normal to the electron gas as a function of the top gate voltage which changes the electron concentration, is shown in Fig. 16.1. The general trend is that the conductance increases with increasing gate voltage (electron density). At the same time, the oscillation amplitude increases while the oscillations are periodic in the gate voltage.

Figure 16.2 shows a measurement in which the magnetic field is swept. It was performed on a two-dimensional electron gas in a GaAs quantum

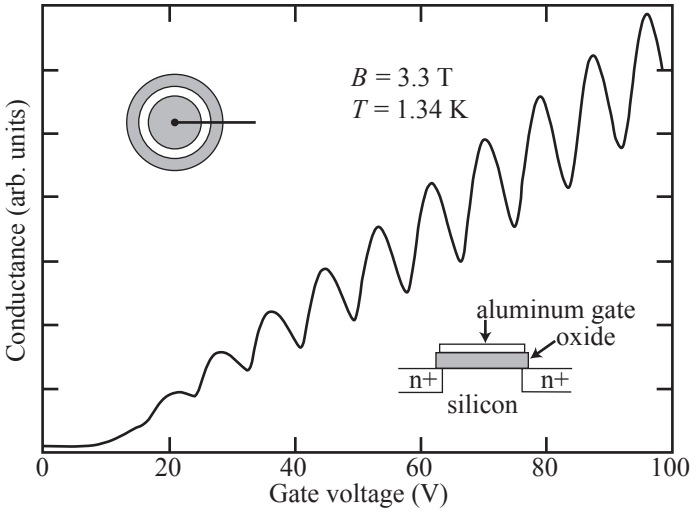


Fig. 16.1 Shubnikov–de Haas oscillations in σ_{xx}, measured on a Si MOS structure in Corbino geometry at a temperature of 1.34 K and in a magnetic field of 3.3 T. (Reprinted with permission from Fowler *et al.*, 1966. Copyright 1966 by the American Physical Society.)

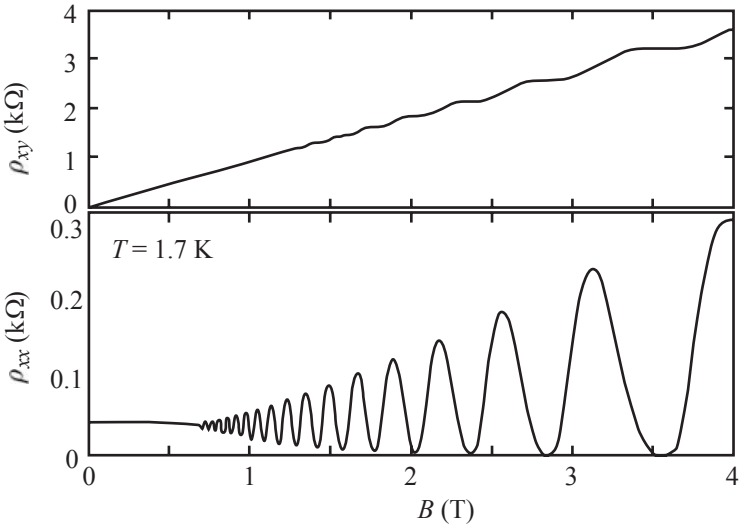


Fig. 16.2 Shubnikov–de Haas oscillations in the longitudinal resistivity (bottom) and the Hall resistivity (top) of a two-dimensional electron gas in a 10 nm wide GaAs quantum well. The measurement was performed at the temperature $T = 1.7$ K.

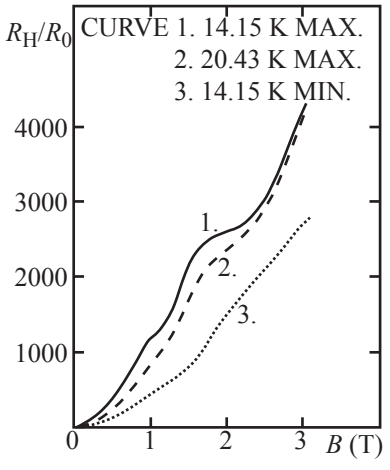


Fig. 16.3 Shubnikov–de Haas oscillations in the Hall coefficient of bismuth, measured at the temperature $T = 14.15$ K. (Shubnikov and de Haas, 1930*a*. Courtesy of the Leiden Institute of Physics.)

well of 10 nm width in which a single subband was occupied. A Hall bar structure allowed the simultaneous measurement of the resistivity tensor components ρ_{xx} and ρ_{xy}. For magnetic fields $B < 0.5$ T, the longitudinal resistivity is almost constant and the Hall resistivity increases linearly, as expected from the Drude model. For larger magnetic fields, both quantities oscillate around the classical magnetoresistivity. The period of the oscillations increases with increasing magnetic field. A detailed analysis shows that the oscillations are periodic in $1/B$.

The effect described above is known from the conductance of metallic samples as the *Shubnikov–de Haas effect*. It was discovered by L. Shubnikov and W.J. de Haas on three-dimensional bismuth samples around 1930 (Shubnikov and de Haas, 1930*a,b,c,d*). Their original measurements are shown in Fig. 16.3.

16.1.1 Electron in a perpendicular magnetic field

In order to understand the origin of the Shubnikov–de Haas effect, we have to quantize the classical cyclotron motion. In a first intuitive step we will do this by applying the Bohr–Sommerfeld quantization scheme. In a second step, we will solve Schrödinger's equation for the electron in a magnetic field.

Bohr–Sommerfeld quantization. We can quantize the classical cyclotron motion by realizing that the electron behaves like a wave. The wave will propagate around a circle and interfere with itself. The self-interference is constructive and a quantized orbit forms if the acquired phase is an integer multiple of 2π, otherwise interference is destructive and no quantum state exists.

We describe the homogeneous magnetic field $\mathbf{B} = (0, 0, B)$ with the

vector potential $\mathbf{A} = (-By/2, Bx/2, 0)$. The strength of the vector potential is $BR_c/2$ along circles of radius R_c around the origin. The radius of the classical cyclotron orbit, R_c, is related to the velocity v of the electron via $R_c = v/\omega_c = mv/eB$. The classical momentum of a particle with charge $-|e|$ in a magnetic field is $\mathbf{p} = m\mathbf{v} - |e|\mathbf{A}$. The term $m\mathbf{v}$ is called the kinetic momentum, $-|e|\mathbf{A}$ is the momentum of the field.

Using the de Broglie relation $m\mathbf{v} = \hbar\mathbf{k}$, we can express the classical cyclotron radius as $R_c = \hbar k/eB$. Like the classical momentum, the quantum phase acquired by the electron during one revolution of length $2\pi R_c$ around the cyclotron orbit consists of two parts. The first, called dynamic phase, is given by

$$\Delta\varphi_d = 2\pi k R_c = 4\pi \frac{\phi}{\phi_0},$$

where we have introduced the magnetic flux through the area πR_c^2 of the circular orbit, given by $\phi = B\pi R_c^2$, and the magnetic flux quantum $\phi_0 = h/e$. The second contribution is the Aharonov–Bohm phase (14.4). It is determined from the integral of the vector potential $\mathbf{A}$ along the circular path taken by the electron, i.e.,

$$\Delta\varphi_{AB} = -\frac{|e|}{\hbar}\oint \mathbf{A}d\mathbf{s} = -\frac{|e|}{\hbar}\frac{BR_c}{2}2\pi R_c = -2\pi\frac{\phi}{\phi_0}.$$

The Bohr–Sommerfeld quantization of the electron motion therefore leads to the condition

$$\Delta\varphi = \Delta\varphi_d + \Delta\varphi_{AB} = 2\pi\frac{\phi}{\phi_0} = 2\pi n,$$

with n being a positive integer. This condition states that the magnetic flux enclosed by the electronic orbit is quantized in units of the flux quantum ϕ_0. As a consequence, the radius of possible cyclotron orbits at a given magnetic field is quantized and given by

$$l_c^{(n)} = \sqrt{2n}\, l_c,$$

where the magnetic length $l_c = \sqrt{\hbar/eB}$ is the characteristic length scale of the cyclotron motion at a given magnetic field. This quantization of the classical cyclotron orbits is visualized in Fig. 16.4. The energy of the electron is quantized according to

$$E_n = \frac{1}{2}m\omega_c^2 {l_c^{(n)}}^2 = \hbar\omega_c n, \tag{16.1}$$

indicating that the cyclotron energy $\hbar\omega_c$ is the relevant energy scale of the problem. The energy levels form a ladder with constant spacing like those of the harmonic oscillator. The energy spectrum of an electron in a magnetic field resembles that of a harmonic oscillator with the characteristic frequency ω_c determined by the magnetic field.

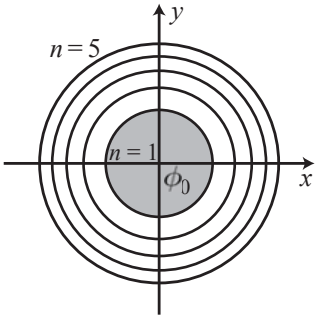


Fig. 16.4 Quantized cyclotron orbits in real space. The smallest orbit encloses one single flux quantum ϕ_0.

Solution of Schrödinger's equation. Although the Bohr–Sommerfeld quantization gives the characteristic length scale l_c and the characteristic energy scale $\hbar\omega_c$ and highlights the effect of the Aharonov–Bohm contribution to the electronic phase, an exact solution of Schrödinger's equation is instructive. The effective mass hamiltonian for a parabolic band reads

$$H = \frac{(\mathbf{p} + |e|\mathbf{A})^2}{2m^\star} + V(z),$$

where $V(z)$ is the confinement potential in the growth direction which may be caused by a heterointerface and remote doping, or by a quantum well potential. For simplicity we choose the vector potential $\mathbf{A} = (-By, 0, 0)$ describing the magnetic field $B = (0, 0, B)$ oriented in the z-direction. This hamiltonian can be separated into

$$H_z = -\frac{\hbar^2}{2m^\star}\frac{\partial^2}{\partial z^2} + V(z),$$

depending only on the z-coordinate, but not on magnetic field, and

$$H_{xy} = \frac{(p_x - |e|B_z y)^2 + p_y^2}{2m^\star},$$

which is independent of the confinement potential, but contains the magnetic field. The eigenvalue problem in the z-direction leads to bound states which are independent of the magnetic field. In a two-dimensional electron gas in the quantum limit, only the lowest of these states will be occupied.

In the plane, the problem is solved using the *Ansatz*

$$\psi(x, y) = e^{ik_x x}\eta(y),$$

leading to the eigenvalue problem

$$\left[\frac{p_y^2}{2m^\star} + \frac{1}{2}m^\star\omega_c^2\left(y - \frac{\hbar k_x}{|e|B_z}\right)^2\right]\eta_{k_x}(y) = E\eta_{k_x}(y), \tag{16.2}$$

where we have introduced the cyclotron frequency $\omega_c = |e|B/m^\star$. This is the equation of a one-dimensional quantum mechanical harmonic oscillator with the k_x-dependent center coordinate

$$y_0 = \frac{\hbar k_x}{|e|B}.$$

As a result, the quantized energy states are given by

$$E_n = \hbar\omega_c\left(n + \frac{1}{2}\right),$$

independent of k_x. Quantum states with different quantum numbers k_x but the same quantum number n are energetically degenerate. All the states of different k_x but the same n form the so-called *Landau-level*. Compared to the quantized energies obtained from the Bohr–Sommerfeld quantization, the exact result requires the replacement $n \rightarrow$

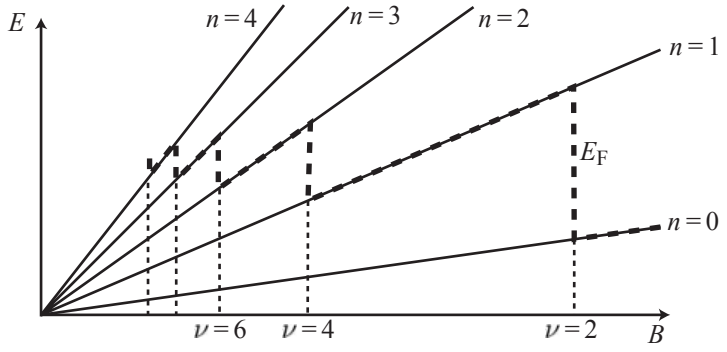


Fig. 16.5 Energy levels for electrons in a magnetic field. The energy of the Landau levels increases linearly with the magnetic field resulting in a fan-like diagram. The slope of each Landau level line depends on the quantum number n. At fixed electron density in the electron gas, the Fermi energy oscillates as a function of the filling factor (dashed line).

$n + 1/2$ in eq. (16.1), i.e., the Bohr–Sommerfeld result is only correct for large quantum numbers n, when the addition of $1/2$ is irrelevant.

The energy of a given Landau level increases linearly with the magnetic field B. This leads to the so-called *Landau fan* which is the energy diagram depicted in Fig. 16.5.

The degeneracy of a Landau level is given by the requirement that the center coordinate $y_0 = \hbar k_x/eB$ has to be within the width W of the structure, i.e., $0 \leq \hbar k_x/eB \leq W$. For a two-dimensional electron gas of length L, the density of k_x-states is $L/2\pi$. As a result, meaningful k_x-values obey the relation $0 \leq k_x L/2\pi \leq eB/hA$, where $A = WL$ is the sample area. The number n_L of allowed k_x states per unit area is therefore

$$n_\mathrm{L} = \frac{|e|B}{h}.$$

If an electron gas has the electron density n_s, the number $\nu = n_\mathrm{s}/n_\mathrm{L}$ tells us how many Landau levels are occupied at a given magnetic field at zero temperature. Therefore, $\nu = hn_\mathrm{s}/|e|B$ is called the filling factor corresponding to the magnetic field B. At a fixed electron density n_s, the Fermi level of the electron gas oscillates as a function of B, i.e., with filling factor ν in a $1/B$-periodic fashion as shown in Fig. 16.5. The spin of the electrons was neglected in the above consideration. If the Zeeman splitting $g^\star \mu_\mathrm{B} B$ is negligible compared to the Landau level splitting $\hbar\omega_\mathrm{c}$, each Landau level hosts $2n_\mathrm{L}$ electrons per unit area and the Fermi energy jumps between Landau levels at even values of ν (see Fig. 16.5).

If we take the Zeeman splitting of electronic levels into account, the Zeeman energy adds to the Landau level energy and we obtain the spectrum

$$E_n^\pm = \hbar\omega_c \left(n + \frac{1}{2}\right) \pm \frac{1}{2} g^\star \mu_B B_z.$$

The factor $g^\star$ can depend strongly on the magnetic field, because it may be renormalized by exchange interaction effects in the two-dimensional electron gas.

In our present model, the density of states in a magnetic field is given

by

$$\mathcal{D}_{\mathrm{2D}}(E,B) = \frac{|e|B}{h} \sum_{n,\sigma=\pm} \delta\left(E - E_n^{(\sigma)}\right).$$

How can we understand the transition from the discrete, strongly degenerate density of states in a magnetic field to the continuous, constant density of states at $B = 0$? At small magnetic fields, the number of occupied Landau levels is large and given by $\nu = E_{\mathrm{F}}/\hbar\omega_{\mathrm{c}}$. The number of states per spin-degenerate Landau level is $2|e|B/h$. The density of occupied states at low magnetic fields is therefore given by

$$n_{\mathrm{s}} = \frac{2|e|B}{h}\frac{E_{\mathrm{F}}}{\hbar\omega_{\mathrm{c}}} = \frac{m^\star}{\pi\hbar^2}E_{\mathrm{F}}.$$

This means that, in the limit of $B \to 0$, the Landau levels decrease their separation such that they eventually form the constant two-dimensional density of states. The energetic broadening of Landau levels to be discussed below contributes to the disappearance of the discrete density of states peaks.

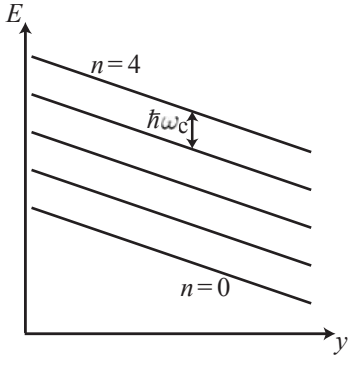


Fig. 16.6 Real space representation of the Landau level spectrum in the presence of an electric field in the y-direction. The slope of the Landau levels is given by $-eE$, and their spacing is $\hbar\omega_{\mathrm{c}}$.

16.1.2 Quantum treatment of E × B-drift

In the classical treatment of electron motion in chapter 10 we considered an electric field in the plane of the electron gas in addition to the perpendicular magnetic field, which led to the so-called **E** × **B**-drift. We will now present the solution of the corresponding quantum mechanical problem. We introduce the electric field $\mathbf{E} = (0, E, 0)$ described by the electrostatic potential $V(y) = |e|Ey$. It is only relevant for the equation of motion in the y-direction and adds to the hamiltonian of the eigenvalue problem in eq. (16.2) giving

$$\left[\frac{p_y^2}{2m^\star} + \frac{1}{2}m^\star\omega_c^2\left(y - \frac{\hbar k_x + m^\star v_{\mathrm{D}}}{|e|B_z}\right)^2 - \frac{\hbar k_x + m^\star v_{\mathrm{D}}}{|e|B_z}|e|E_y + \frac{1}{2}m^\star v_{\mathrm{D}}^2\right]\eta_{k_x}(y) = E\eta_{k_x}(y). \quad (16.3)$$

Here, $v_{\mathrm{D}} = E/B$ is the classical drift velocity of the electrons which corresponds to a quantum mechanical group velocity for electrons as we will show below. The above equation is again solved with the harmonic oscillator eigenfunctions, but the center coordinate is modified to

$$\tilde{y}_0 = \frac{\hbar k_x + m^\star v_{\mathrm{D}}}{|e|B},$$

and the wave function of state (n, k_x) is

$$\psi_{nk_x}(x,y) = \frac{1}{\sqrt{2^n n!\pi^{1/2}}}H_n[(y-\tilde{y}_0)/l_{\mathrm{c}}]e^{-(y-\tilde{y}_0)^2/2l_{\mathrm{c}}^2}\frac{1}{\sqrt{L}}e^{ik_x x}, \quad (16.4)$$

where $l_c = \sqrt{\hbar/eB}$ is the magnetic length introduced before. The energy of state (n, k_x) is given by

$$E_n(k_x) = \hbar\omega_c \left(n + \frac{1}{2}\right) + |e|E_y\tilde{y}_0 + \frac{1}{2}m^\star v_D^2. \tag{16.5}$$

This result can be interpreted as follows: compared to the energy spectrum without an electric field, the cyclotron states have additional potential energy $|e|E_y\tilde{y}_0$ resulting from the position of the wave function in the electrostatic potential, and the additional kinetic energy $m^\star v_D^2/2$ resulting from drift motion in the x-direction. The previous degeneracy of Landau levels has been completely lifted by the electric field. Owing to the close relation between k_x and $\tilde{y}_0$, the energy spectrum can be represented in real space as a tilted ladder of states as depicted in Fig. 16.6.

The group velocity of electrons in a certain Landau level is given by

$$\frac{1}{\hbar}\frac{\partial E_n(k_x)}{\partial k_x} = v_D$$

corresponding to the classical result. The expectation value of the momentum in the y-direction is zero. This has an interesting implication for the tensor of the conductivity. It means that the longitudinal conductivity $\sigma_{xx} = 0$ (along the direction of $\mathbf{E}$), whereas $\sigma_{xy} = |e|n_s v_D/E = |e|n_s/B$. Tensor inversion gives $\rho_{xx} = 0$ and $\rho_{xy} = B/|e|n_s$, i.e., the classical Hall resistance. We see in this example that in a magnetic field, σ_{xx} and ρ_{xx} can be zero at the same time, if the Hall conductivity σ_{xy} and the Hall resistivity ρ_{xy} remain finite.

Of course, this view of the conductivity in a magnetic field neglects completely the effect of scattering which is crucial for a proper understanding of electrical resistance in a magnetic field. This will be considered in the following discussion of the Shubnikov–de Haas effect and the quantum Hall effect.

16.1.3 Landau level broadening by scattering

Spatial potential fluctuations as they are at low temperatures, for example created by the random arrangement of charged dopants, lift the degeneracy of the states of a Landau level. This effect can be seen as the influence of scattering at potential fluctuations limiting the lifetime of an electron in a certain quantum state. Possible scattering processes include intra- and inter-Landau-level scattering. As a result, the ideal delta-shaped density of states peaks are broadened as depicted in Fig. 16.7.

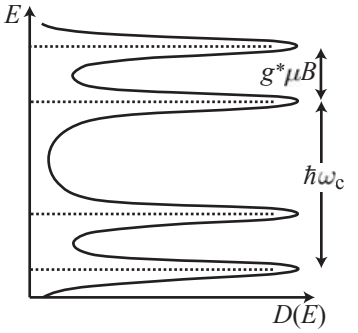


Fig. 16.7 Density of states for electrons in a magnetic field. The Landau levels are broadened by scattering at spatial potential fluctuations of the electron gas. The density of states peaks have an energetic separation of $\hbar\omega_c$. In addition, the electron spin leads to a Zeeman splitting of Landau levels.

Short range scattering potentials. In the case of short-range scattering potentials with a mean scattering rate between states of $1/\tau_q$, the time–energy uncertainty relation suggests an energetic Landau level broadening by $\hbar/\tau_q$. At low magnetic fields, we may estimate the scat-

tering rate by

$$\frac{\hbar}{\tau_{\rm q}(E)} = n_i \frac{m^\star}{2\pi\hbar^2} \int_0^{2\pi} d\varphi \left\langle \left| v^{(i)}(\mathbf{q}) \right|^2 \right\rangle_{\rm imp}. \tag{16.6}$$

Note that the factor in front of the integral is half the two-dimensional density of states (half, because scattering conserves the spin, so only half the density of states is available to scatter into). In contrast to the Drude scattering time in eq. (10.50) in which backscattering has an enhanced weight, for the lifetime broadening of **k**-states the scattering angle is not relevant.

The determination of the lifetime of quantum states at arbitrary magnetic fields is problematic due to an interdependence of the scattering rate and the density of states. A large density of states at the Fermi energy leads to an enhanced scattering rate because, according to first order perturbation theory [cf., eq. (16.6)],

$$\frac{1}{\tau_{\rm q}(E)} = \frac{2\pi}{\hbar} n_i \overline{v^2} \mathcal{D}(E), \tag{16.7}$$

where $\overline{v^2}$ is an angle and impurity ensemble averaged squared scattering matrix element and $\mathcal{D}(E)$ is the density of states at energy E (of one particular spin species). On the other hand, the peak density of states depends on the scattering rate $\tau_{\rm q}^{-1}$. A large scattering rate results in strong broadening of Landau levels, but this decreases the peak density of states, because the integrated density of states must remain equal to the Landau level degeneracy $n_{\rm L}$.

In order to illustrate this interdependency, let us neglect inter-Landau-level scattering and assume that the lifetime broadening leads to a lorentzian density of states

$$\mathcal{D}(E) = \frac{n_{\rm L}}{\pi} \frac{\hbar/2\tau_{\rm q}}{(E-E_0)^2 + (\hbar/2\tau_{\rm q})^2} \tag{16.8}$$

for a single Landau level with full width at half maximum $\hbar/\tau_{\rm q}$. Inserting this expression into eq. (16.7) gives

$$1 = n_i \overline{v^2} n_{\rm L} \frac{1}{(E-E_0)^2 + (\hbar/2\tau_{\rm q})^2}.$$

It follows that

$$(E-E_0)^2 + (\hbar/2\tau_{\rm q})^2 = n_i \overline{v^2} n_{\rm L}$$

and therefore

$$\hbar/2\tau_{\rm q} = \sqrt{n_i \overline{v^2} n_{\rm L} - (E-E_0)^2}.$$

Inserting these two expressions into eq. (16.8) we find the elliptic density of states (see Fig. 16.8)

$$\mathcal{D}(E) = \frac{n_{\rm L}}{\pi\Gamma} \sqrt{1 - \left(\frac{E-E_0}{\Gamma}\right)^2}$$

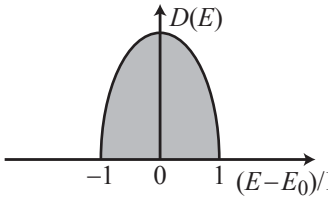


Fig. 16.8 Elliptic density of states of a Landau level as obtained from the self-consistent Born approximation.

for a single Landau level with the characteristic Landau level broadening

$$\Gamma = \sqrt{\frac{1}{2\pi}\hbar\omega_c \frac{\hbar}{\tau_0}}, \tag{16.9}$$

and $\tau_0^{-1} = (\pi/\hbar) n_i \overline{v^2} \mathcal{D}_{2D}$ is the zero magnetic field quantum scattering rate. Essentially the same result for the density of states and the level broadening has been obtained by Ando for short range scatterers within the self-consistent Born approximation (Ando *et al.*, 1982). Within this model, the Landau level broadening increases proportional to $\sqrt{B}$ and Γ is independent of the Landau level quantum number n. This model illustrates the interdependence of the scattering rate and the density of states, but leads to an unphysical behavior of the density of states in the tails of the Landau level. Another approach (Gerhardts, 1975; Gerhardts, 1976) has led to gaussian broadening of Landau levels (see Fig. 16.9)

$$\mathcal{D}(E) = n_L \sqrt{\frac{2}{\pi\Gamma^2}} \exp\left(-2\frac{(E-E_0)^2}{\Gamma^2}\right).$$

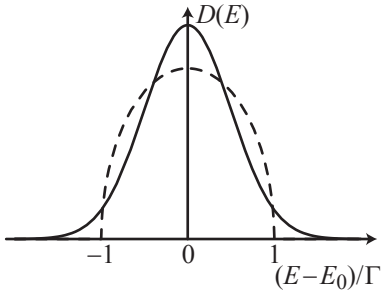


Fig. 16.9 Landau level density of states with gaussian broadening (solid line) and with semielliptic broadening (dashed line).

Long-range scattering potentials. An intuitive picture of Landau level broadening can also be found in the case of long-range scattering potentials. The energy of the Landau levels follows the local variations of the disorder potential adiabatically and the density of states is broadened like the distribution function of the potential fluctuations. Again, the broadening is independent of the Landau level quantum number n.

Low magnetic field oscillations of the density of states. We will now develop a description of the density of states in a magnetic field which is suitable for further calculations at low magnetic fields, where many Landau levels are occupied ($\hbar\omega_c \ll E_F$). If an individual Landau level with quantum number n has the density of states $n_L L_n[E - \hbar\omega_c(n + 1/2)]$ the total spin-degenerate density of states can be written as

$$\mathcal{D}_{2D}(E, B) = 2n_L \sum_n L_n \left(E - \hbar\omega_c(n + 1/2)\right).$$

Figure 16.10 shows such a density of states, where L_n has been assumed to be a lorentzian with $\hbar/\tau_q$ being the full width at half maximum independent of n. The sum over Landau levels leads to an oscillatory behavior as a function of energy. If we assume that the Landau level density of states is independent of the Landau level quantum number n, i.e., $L_n(E) \equiv L(E)$, the expression for the density of states can be rewritten using Poisson's summation formula

$$\sum_{n=0}^{\infty} f(n + 1/2) = \int_0^{\infty} f(x)dx + 2\sum_{s=1}^{\infty}(-1)^s \int_0^{\infty} f(x)\cos(2\pi x s)dx.$$

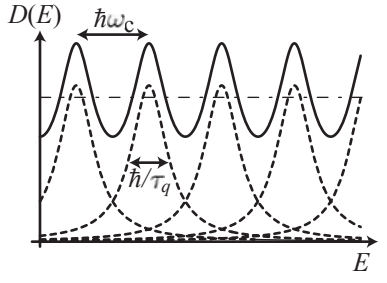


Fig. 16.10 Density of states of a two-dimensional electron gas as a function of energy at finite magnetic fields (solid line). It is composed of the sum of a number of individual Landau levels with separation $\hbar\omega_c$ and width $\hbar/\tau_q$ (dashed lines). The zero magnetic field density of states is the dash-dotted line.

We have

$$\frac{1}{2n_{\rm L}}\mathcal{D}_{\rm 2D}(E,B) = \sum_n L\left(E-\hbar\omega_{\rm c}(n+1/2)\right)$$
$$= \int_0^\infty L(E-\hbar\omega_{\rm c}x)dx + 2\sum_{s=1}^{\infty}(-1)^s\int_0^\infty L(E-\hbar\omega_{\rm c}x)\cos(2\pi xs)dx$$
$$= \frac{1}{\hbar\omega_{\rm c}}\left[\int_{-\infty}^{E} L(\xi)d\xi + 2\sum_{s=1}^{\infty}(-1)^s\int_{-\infty}^{E} L(\xi)\cos(2\pi s(E-\xi)/\hbar\omega_{\rm c})d\xi\right].$$

For energies $E \gg \hbar/\tau$ we can replace the upper integration limit by $+\infty$ which leads to

$$\mathcal{D}_{\rm 2D}(E,B) = \frac{m^\star}{\pi\hbar^2}\left[1+\frac{\Delta\mathcal{D}}{\mathcal{D}}\right]. \tag{16.10}$$

The first term in the square brackets is the contribution of the constant two-dimensional density of states at zero magnetic field. The second term is a sum over oscillatory contributions to the density of states given by

$$\begin{aligned}\frac{\Delta\mathcal{D}}{\mathcal{D}} &= 2\sum_{s=1}^{\infty}(-1)^s\int_{-\infty}^{\infty} L(\xi)\cos(2\pi s(E-\xi)/\hbar\omega_{\rm c})d\xi \\ &= 2\sum_{s=1}^{\infty}(-1)^s\cos(2\pi sE/\hbar\omega_{\rm c})\int_{-\infty}^{\infty} L(\xi)\cos(2\pi s\xi/\hbar\omega_{\rm c})d\xi \\ &\quad +2\sum_{s=1}^{\infty}(-1)^s\sin(2\pi sE/\hbar\omega_{\rm c})\int_{-\infty}^{\infty} L(\xi)\sin(2\pi s\xi/\hbar\omega_{\rm c}).\end{aligned}$$

If the Landau levels are symmetrically broadened, i.e., if $L(\xi) = L(-\xi)$, the sin-terms vanish for symmetry reasons and we obtain

$$\frac{\Delta\mathcal{D}}{\mathcal{D}} = 2\sum_{s=1}^{\infty}(-1)^s\tilde{L}\left(\frac{2\pi s}{\hbar\omega_{\rm c}}\right)\cos(2\pi sE/(\hbar\omega_{\rm c})), \tag{16.11}$$

where $\tilde{L}(x)$ is the Fourier cosine transform of $L(E)$. For example, the lorentzian density of states in eq. (16.8) has the Fourier cosine transform

$$\tilde{L}(2\pi s/(\hbar\omega_{\rm c})) = e^{-\pi s/(\omega_{\rm c}\tau_{\rm q})}.$$

As a result, the amplitude of the density of states modulation increases exponentially with increasing magnetic field, whereas it is exponentially suppressed for decreasing B. This exponential factor that accounts for the effect of the finite Landau level width on the density of states oscillations is known as the *Dingle factor* (Dingle, 1952). It is the reason that at small magnetic fields, $\omega_{\rm c}\tau \ll 1$, it is sufficient to consider the first term with $s = 1$ and the density of states has the harmonic variation

$$\frac{\Delta\mathcal{D}}{\mathcal{D}} = -2\tilde{L}\left(\frac{2\pi}{\hbar\omega_{\rm c}}\right)\cos\left(2\pi\frac{E}{\hbar\omega_{\rm c}}\right).$$

At a constant energy, e.g., at the Fermi energy, the density of states varies periodically in $1/B$. The amplitude of the modulation increases with increasing magnetic field as a result of the increase of $\tilde{L}\,(2\pi/\hbar\omega_\mathrm{c})$.

From the above discussion we can conclude that the modulation of the density of states can be observed more easily in samples with high mobilities than in those with low mobilities. High mobility samples show effects originating from the oscillatory density of states, such as Shubnikov–de Haas oscillations, even at lower magnetic fields.

16.1.4 Magnetocapacitance measurements

In chapter 9 we calculated the capacitance between the two-dimensional electron gas in a heterostructure and a top gate [cf., eq. (9.1)]:

$$\frac{1}{C/A} = \frac{s+d}{\varepsilon\varepsilon_0} + \frac{1}{e^2}\frac{dE_0(n_\mathrm{s})}{dn_\mathrm{s}} + \frac{1}{e^2}\frac{dE_\mathrm{F}(n_\mathrm{s})}{dn_\mathrm{s}}.$$

The first term describes the geometric capacitance of a parallel plate capacitor with separation $s+d$ of the plates, the second term is proportional to the distance from the heterointerface to the center of mass of the wave function, and the third term contains the inverse, so-called *thermodynamic density of states*. We can therefore write

$$\frac{1}{C/A} = \frac{s+d+\gamma\langle z\rangle}{\varepsilon\varepsilon_0} + \frac{1}{e^2\mathcal{D}_\mathrm{2D}^\mathrm{(th)}(E_\mathrm{F},B)},$$

where $\langle z\rangle$ is the center of mass of the wave function, γ is a numerical constant, and $\mathcal{D}_\mathrm{2D}^\mathrm{(th)}(E_\mathrm{F},B) = dn_\mathrm{s}/dE_\mathrm{F}$ is the thermodynamic density of states at the Fermi level. The geometric contribution to the capacitance and $\langle z\rangle$ do not depend on the magnetic field. The magnetic-field-dependent thermodynamic density of states can therefore be directly determined from a measurement of the capacitance.

In order to illustrate the meaning of the term thermodynamic density of states, we calculate

$$\frac{dn_\mathrm{s}}{dE_\mathrm{F}} = \frac{d}{dE_\mathrm{F}}\int_0^\infty dE\mathcal{D}(E,B)f^{(0)}(E) = \int_0^\infty dE\mathcal{D}(E,B)\frac{df^{(0)}(E)}{dE_\mathrm{F}}.$$

The derivative of the Fermi–Dirac distribution with respect to the Fermi energy is the same as the negative derivative with respect to the energy appearing in the equations for the Drude conductivity in eqs (10.38) and (10.39). For the density of states we can use eqs (16.10) and (16.11). As a result we obtain integrals of the form

$$\begin{aligned}\int_{-\infty}^{\infty} dE\cos\left(2\pi s\frac{E}{\hbar\omega_\mathrm{c}}\right)\left(-\frac{\partial f^{(0)}(E)}{\partial E}\right)&\\ = \frac{1}{2}\cos\left(\frac{2\pi sE_\mathrm{F}}{\hbar\omega_\mathrm{c}}\right)\int_{-\infty}^{\infty} d\eta\frac{\cos\left(\frac{4\pi sk_\mathrm{B}T}{\hbar\omega_\mathrm{c}}\eta\right)}{\cosh^2\eta}&\\ = -\cos\left(\frac{2\pi sE_\mathrm{F}}{\hbar\omega_\mathrm{c}}\right)\frac{X_s}{\sinh X_s},&\quad (16.12)\end{aligned}$$

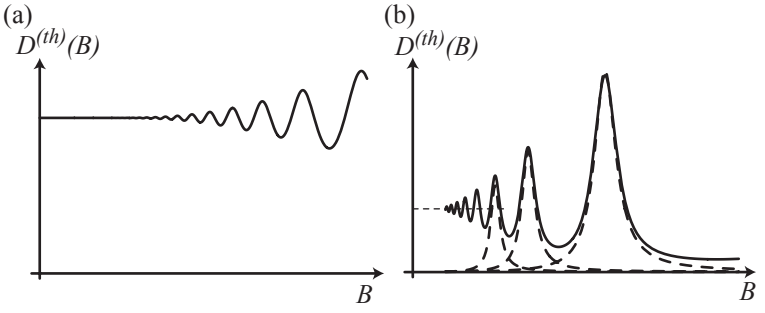


Fig. 16.11 (a) Thermodynamic density of states at constant energy for low magnetic fields. The oscillations are periodic in $1/B$. Damping at low fields is due to the Dingle factor and to the temperature-dependent damping term in eq. (16.13). (b) Thermodynamic density of states at constant energy for high magnetic fields (solid line). The oscillations are periodic in $1/B$. Terms in eq. (16.13) with $s > 1$ are relevant. The thick dashed lines indicate the contributions of the individual Landau levels. The thin dashed line at low fields is the constant zero field density of states.

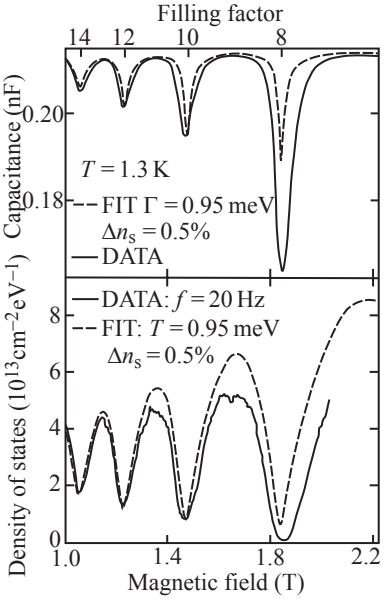


Fig. 16.12 (a) Measured and calculated capacitance of a GaAs/AlGaAs heterostructure in a magnetic field applied normal to the plane of the two-dimensional electron gas. (b) The extracted density of states (measured: solid line, calculated: dashed line). The calculation is based on Landau levels with gaussian broadening. (Reprinted with permission from Smith *et al.*, 1985. Copyright 1985 by the American Physical Society.)

where $X_s = 2\pi^2 s k_\mathrm{B} T/\hbar\omega_\mathrm{c}$. The lower bound of the integration can be set to $-\infty$ as long as $k_\mathrm{B}T \ll E_\mathrm{F}$. The density of states oscillations are smeared by the derivative of the Fermi–Dirac distribution function. At a given magnetic field, increasing temperature T reduces the oscillation amplitude by the factor $X_s/\sinh X_s = X_s \mathrm{csch} X_s$ and the thermodynamic density of states becomes

$$\frac{dn_\mathrm{s}}{dE_\mathrm{F}} = \frac{m^\star}{\pi\hbar^2}\left[1 + 2\sum_{s=1}^{\infty}(-1)^s \tilde{L}\left(\frac{2\pi s}{\hbar\omega_\mathrm{c}}\right)\frac{X_s}{\sinh X_s}\cos\left(\frac{2\pi s E_\mathrm{F}}{\hbar\omega_\mathrm{c}}\right)\right]. \tag{16.13}$$

A plot of the low-magnetic-field thermodynamic density of states is shown in Fig. 16.11(a). At higher magnetic fields, the $1/B$-periodicity is more visible [see Fig. 16.11(b)]. At the same time, the density of states does not exhibit the damped oscillator behavior characteristic for $s = 1$, but higher order terms with $s > 1$ are of significant importance.

The first measurements of this type were on two-dimensional electron gases in silicon MOSFETs performed by Kaplit and Zemel, 1968. An experimental difficulty of this method is that the electron gas has to be charged with a current flowing through the electron gas magnetoresistance. The problem can be described by a distributed *R-C* circuit. Resistive effects can be minimized if the frequency of the measurement is kept as small as possible. Figure 16.12 shows the measured magnetocapacitance of a GaAs/AlGaAs heterostructure and the extracted density of states in comparison to a calculation based on Landau levels with gaussian broadening (Smith *et al.*, 1985).

16.1.5 Oscillatory magnetoresistance and Hall resistance

The actual calculation of the longitudinal and transverse conductivities in the low-magnetic-field Shubnikov–de Haas regime are rather complex. Calculations are, for example, found in Ando *et al.* 1982; Isihara and Smrčka 1986; Laikhtman and Altshuler 1994. Their results can be de-

rived by using plausibility arguments based on the incorporation of the oscillatory density of states given by eq. (16.11) into the Drude conductivity tensor components in eqs (10.36) and (10.37). We assume that the Landau levels have a lorenzian density of states and keep only the lowest order $s = 1$ contribution in eq. (16.11). In this low-magnetic-field approximation we have

$$\frac{\Delta\mathcal{D}}{\mathcal{D}} = -2\exp\left(-\frac{\pi}{\omega_{\rm c}\tau_q}\right)\cos(2\pi E/(\hbar\omega_{\rm c})),$$

and we can treat this quantity as a small perturbation. We can argue that the energy-dependent scattering rate is proportional to the density of states according to eq. (16.7) which leads to

$$\frac{1}{\tau(E)} = \frac{1}{\tau_0(E)}\left(1+\frac{\Delta\mathcal{D}}{\mathcal{D}}\right) \Rightarrow \tau(E) = \tau_0(E)\left(1-\frac{\Delta\mathcal{D}}{\mathcal{D}}\right),$$

where $\tau_0^{-1}(E)$ is the zero-magnetic-field scattering rate. Inserting this expansion into the Drude result in eqs (10.36) and (10.37), and keeping only terms linear in $\Delta\mathcal{D}/\mathcal{D}$ gives

$$\begin{aligned}\sigma_{xx}(E) &= \frac{ne^2\tau_0/m^\star}{1+\omega_{\rm c}^2\tau_0^2}\left[1-\frac{1-\omega_{\rm c}^2\tau_0^2}{1+\omega_{\rm c}^2\tau_0^2}\frac{\Delta\mathcal{D}(E)}{\mathcal{D}}\right]\\ \sigma_{xy}(E) &= \frac{ne^2\omega_{\rm c}\tau_0^2/m^\star}{1+\omega_{\rm c}^2\tau_0^2}\left[1-\frac{2}{1+\omega_{\rm c}^2\tau_0^2}\frac{\Delta\mathcal{D}(E)}{\mathcal{D}}\right].\end{aligned}$$

Thermal averaging with the derivative of the Fermi–Dirac distribution function according to eq. (10.36), where we neglect the energy dependence of τ_0, leads to integrals of the form (16.12) and produces, in agreement with Laikhtman and Altshuler, 1994,

$$\sigma_{xx}(B,T) = \frac{ne^2\tau_0/m^\star}{1+\omega_{\rm c}^2\tau_0^2} \left[1+2\frac{1-\omega_{\rm c}^2\tau_0^2}{1+\omega_{\rm c}^2\tau_0^2}e^{-\pi/\omega_{\rm c}\tau_q}\frac{2\pi^2k_{\rm B}T/\hbar\omega_{\rm c}}{\sinh(2\pi^2k_{\rm B}T/\hbar\omega_{\rm c})}\cos\left(2\pi\frac{hn}{2eB}\right)\right] \quad (16.14)$$

$$\sigma_{xy}(B,T) = \frac{ne^2\omega_{\rm c}\tau_0^2/m^\star}{1+\omega_{\rm c}^2\tau_0^2} \left[1+\frac{4}{1+\omega_{\rm c}^2\tau_0^2}e^{-\pi/\omega_{\rm c}\tau_q}\frac{2\pi^2k_{\rm B}T/\hbar\omega_{\rm c}}{\sinh(2\pi^2k_{\rm B}T/\hbar\omega_{\rm c})}\cos\left(2\pi\frac{hn}{2eB}\right)\right]. \quad (16.15)$$

The relation $E_{\rm F}/\hbar\omega_{\rm c} = hn/2|e|B$ was used here in order to rewrite the argument of the oscillating factor. We emphasize here that τ_0 has to be interpreted as a zero-magnetic-field transport scattering time, whereas τ_q is the lifetime of the quantum states. For long-range scattering potentials, for which the above formulae are relevant, the ratio τ_0/τ_q can be of the order of 10 or more. For short-range scattering potentials the ratio $\tau_0/\tau_q \approx 1$. In this case, the prefactors in front of the oscillating

quantum correction may take a form different from the one given here (see Ando *et al.*, 1982; Isihara and Smrčka, 1986).

The corresponding result for ρ_{xx} is obtained from eqs (16.14) and (16.15) by tensor inversion, where again only terms linear in $\Delta\mathcal{D}/\mathcal{D}$ are kept,

$$\rho_{xx}(B,T) = \frac{m^\star}{ne^2\tau_0} \left[1 - 2e^{-\pi/\omega_c\tau_q} \frac{2\pi^2 k_B T/\hbar\omega_c}{\sinh(2\pi^2 k_B T/\hbar\omega_c)} \cos\left(2\pi\frac{hn}{2eB}\right)\right]. \quad (16.16)$$

The interpretation of this result is straightforward. The prefactor is the magnetic-field-independent classical Drude resistivity around which the magnetoresistance oscillates in a 1/B-periodic fashion. The term in square brackets corresponds to the low-magnetic-field thermodynamic density of states for Landau levels with lorentzian broadening plotted in Fig. 16.11(a). Minima in the longitudinal resistivity arise as a result of minima in the density of states at the Fermi energy. The oscillatory magnetoresistivity in Fig. 16.2 therefore reflects at low magnetic fields the density of states at the Fermi energy. The exponential Dingle factor accounts for the finite lifetime broadening of the Landau levels. The temperature-dependent factor reduces the amplitude of the oscillations as a result of energy averaging over $k_B T$ around the Fermi energy. The Fermi energy itself appears as a constant here, although we know that it oscillates with magnetic field (see Fig. 16.5). This approximation is only justified at low magnetic fields, where $\Delta\mathcal{D}/\mathcal{D}$ is so small that these oscillations can be neglected.

The resistivity component ρ_{xy} resulting from eqs (16.14) and (16.15) does not have oscillatory components in this approximation valid for long-range scattering potentials. This may be different in the case of short-range scatterers (see Ando *et al.*, 1982; Isihara and Smrčka, 1986), where oscillating contributions to ρ_{xy} appear in lowest order. An experimental investigation of the latter can be found in Coleridge *et al.*, 1989.

Electron density determination. From measurements of the Shubnikov–de Haas oscillations the density of a two-dimensional electron gas can be determined. Minima of the magnetoresistance in eq. (16.16) occur for $hn/2|e|B = i + 1/2$, where i is an integer number. As shown in Fig. 16.13, plotting the measured values of $B_i^{-1} = 2|e|i/hn$ vs. the index i gives a straight line with the slope given by $2|e|/hn$. Alternatively, this slope can be extracted from the separation of neighboring minima in $1/B$ according to

$$\Delta\left(\frac{1}{B}\right) = \frac{1}{B_{i+1}} - \frac{1}{B_i} = \frac{2|e|}{hn} = 0.48 \times 10^{15} \frac{\mathrm{m}^{-2}}{\mathrm{T}} \frac{1}{n}.$$

Determination of τ_q. The quantum lifetime τ_q can be determined from Shubnikov–de Haas oscillations at sufficiently low temperatures,

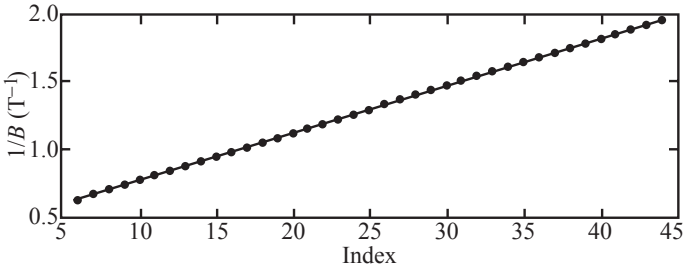


Fig. 16.13 Electron density determination from Shubnikov–de Haas measurements. The indices of magnetoresistance minima are plotted on the horizontal axis, the corresponding values of $1/B$ are plotted on the vertical axis. The slope of the line $2|e|/hn$ determines the electron density.

where the τ_q-dependent Dingle term dominates over the thermal amplitude reduction. According to eq. (16.16) the envelope of the oscillations is given by

$$\frac{\Delta\rho_{xx}}{\bar{\rho}_{xx}} = \pm 2e^{-\pi/(\omega_c\tau_q)} \frac{2\pi^2 k_B T/(\hbar\omega_c)}{\sinh 2\pi^2 k_B T/(\hbar\omega_c)} := \pm 2e^{-\pi/(\omega_c\tau_q)} f(B,T). \tag{16.17}$$

Plotting the quantity

$$\ln\left(\frac{\Delta\rho_{xx}}{\bar{\rho}_{xx}}\frac{1}{f(B,T)}\right) = -\frac{\pi m^\star}{|e|\tau_q}\frac{1}{B} + \text{const.}$$

versus $1/B$ one can extract τ_q from the slope of the resulting straight line. The function $f(B,T)$ is depicted in Fig. 16.14.

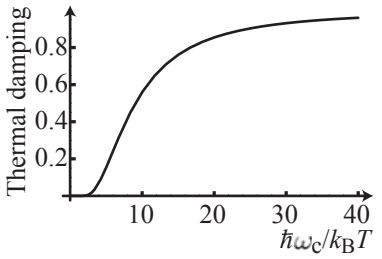


Fig. 16.14 Temperature-dependent factor $2\pi^2 k_B T/\hbar\omega_c \sinh[2\pi^2 k_B T/\hbar\omega_c]$ plotted versus $\hbar\omega_c/(k_B T)$.

Effective mass determination. From measurements of the temperature dependence of the Shubnikov–de Haas oscillations the factor $f(B,T)$ can be measured. The only parameter which is unknown in this parameter is the effective mass $m^\star$ appearing in ω_c. The mass can therefore be determined by comparing the experimentally determined function $f(B,T)$ with plots of this function for different masses.

16.2 Electron localization at high magnetic fields

After the discussion of the low-magnetic-field behavior of the magnetoresistance ρ_{xx} we now return to a discussion of the nature of the states at high magnetic fields. In the discussion in section 16.1.3 the influence of impurity scattering on the density of states was taken into account only in the lowest order of scattering theory. Furthermore, we have completely neglected the fact that screening of the impurity potentials depends on the density of states at the Fermi energy. While these approximations are reasonable at low magnetic fields, where $k_B T, \hbar/\tau_q \leq \hbar\omega_c \ll E_F$, in the high field limit localization of electrons due to an interplay between the random background potential and interactions takes place.

The localization of states manifests itself experimentally in a vanishing magnetoconductivity σ_{xx} when the Fermi energy lies in the vicinity

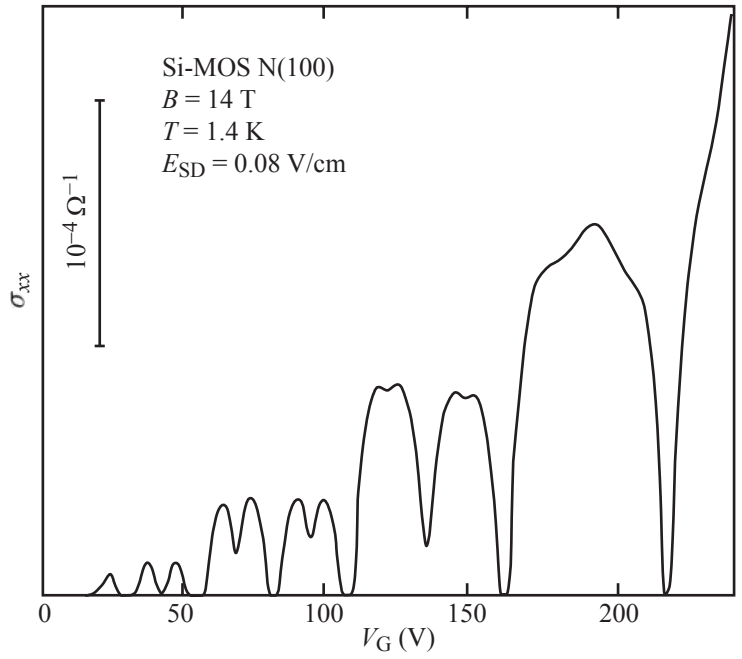


Fig. 16.15 Measurement of σ_{xx} in a two-dimensional electron gas induced in a Si-MOS structure by the application of a gate voltage V_G at $T = 1.4\,K$, $B = 14\,T$ and for a source–drain electric field $E_{SD} = 0.08\,V/cm$. In certain regions of V_G, σ_{xx} vanishes. (Reprinted from Kawaji and Wakabayashi, 1976 with permission from Elsevier.)

of a density of states minimum between adjacent Landau levels. This effect was already known several years before the discovery of the quantum Hall effect. Measurements of σ_{xx} performed at a constant magnetic field of 14 T as a function of the electron density tuned via the top gate voltage are shown in Fig. 16.15. They were performed on a Corbino geometry [cf., eq. (10.22)] fabricated on silicon MOS wafers by Kawaji and Wakabayashi. It can be clearly seen that σ_{xx} vanishes at certain gate voltages V_G. The peaks of σ_{xx} between these zeros correspond to peaks in the density of states at the Fermi energy. The increasing peak height with increasing electron density (increasing V_G) can be qualitatively understood with the following argument. The conductance is given by the Einstein relation (10.54) and the diffusion constant at high magnetic field by eq. (10.56). Because in quantizing magnetic fields the classical cyclotron radius R_c corresponds to the characteristic length scale of the highest occupied Landau level (index N) given by $\sqrt{(2N+1)l_c^2}$, the diffusion constant becomes $D = (2N+1)h/2|e|B\tau$. The peak density of states is, e.g., in the case of a lorentzian given by $\mathcal{D}_{peak} = 2\tau|e|B/\pi h\hbar$ [cf., eq. (16.8)] and therefore the peak conductivity becomes $\sigma_{xx}^{peak} \propto e^2/h(N+1/2)$. With increasing V_G the quantum number N of the highest occupied Landau level increases one by one and therefore the height of the peaks in σ_{xx} increases.

It turns out that in the gate voltage regions, where σ_{xx} vanishes, the Hall conductivity σ_{xy} remains finite. As a consequence of tensor inversion this implies that $\rho_{xx} = 0$ where $\sigma_{xx} = 0$. This result is not very intuitive. We have argued above that localization of states leads

to a vanishing conductivity characteristic for an insulator. However, an insulator (at zero magnetic field) would have $\rho_{xx} \to \infty$. In turn, we would call a material with $\rho_{xx} = 0$ a perfect conductor, and expect (at zero field) an infinite conductivity ($\sigma_{xx} \to \infty$). We see that our intuition is misleading when it comes to high magnetic fields, where the tensor inversion implies $\rho_{xx} = 0 = \sigma_{xx}$, if ρ_{xy} and σ_{xy} remain finite.

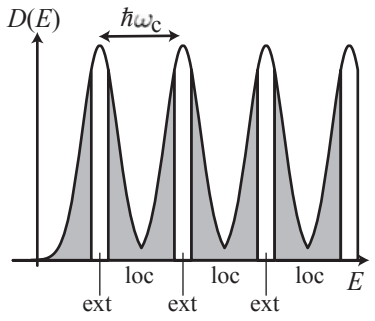


Fig. 16.16 Schematic representation of the density of states $\mathcal{D}(E)$ of Landau levels broadened by disorder as a function of energy E. In the gray shaded regions labeled 'loc', the states are localized, whereas in the white regions labeled 'ext' the states are extended. A mobility edge arises at the transition between extended and localized states.

Localization by spatial potential fluctuations. As was mentioned above, the nature of quantum states is strongly affected by a spatially varying potential. If interactions are neglected, the result is the so-called Anderson localization in a magnetic field. Figure 16.17 shows the modulus of wave functions in the lowest Landau level calculated numerically. A statistical distribution of δ-scatterers was assumed in the plane. We can see that the states in the tails of the broadened density of states are strongly localized [Fig. 16.17(a), (b), (e)]. States in the center of the peaked density of states are extended [Fig. 16.17(c), (d)]. This situation is schematically depicted in the density of states diagram in Fig. 16.16.

If the Fermi energy lies in the tails of the density of states distribution we would expect an exponentially suppressed conductivity, i.e., $\sigma_{xx} \to 0$ for temperature $T \to 0$. If the Fermi energy lies close to a maximum of the density of states, a finite conductivity results. The energies in Fig. 16.16 at which the nature of the states changes from localized to extended or vice versa, define a so-called *mobility edge*.

Localization of electrons also arises in the case of long-range potential fluctuations, if the classical cyclotron radius becomes small compared to

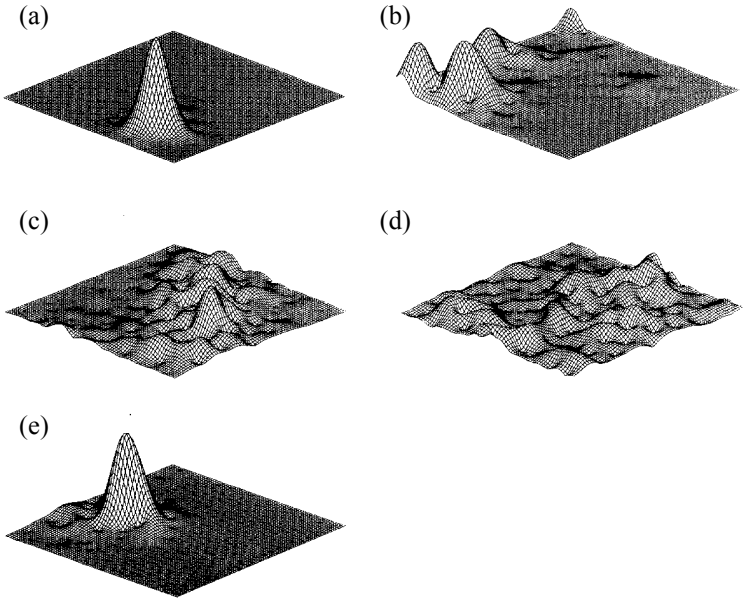


Fig. 16.17 Modulus of normalized wave functions corresponding to different energies in the lowest Landau level calculated for a random distribution of δ-scatterers. Wave functions in (a) and (b) are energetically well below the maximum of the density of states, (c) and (d) are near the density of states maximum, (e) is above the maximum (Aoki, 1977).

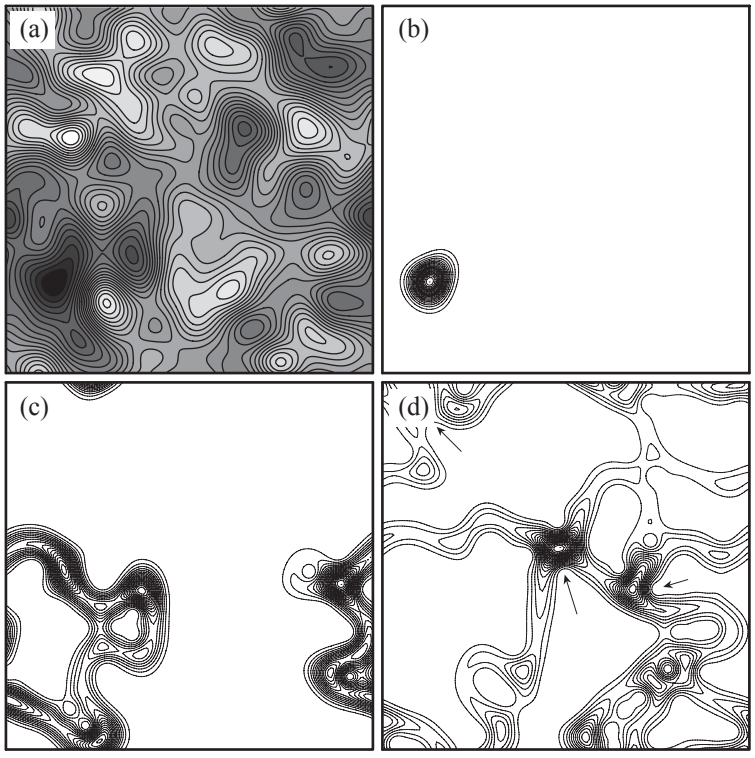


Fig. 16.18 (a) Long-range statistical potential created from a superposition of randomly placed gaussian functions of width $b = 2l_c$. Regions of low potential are dark, those of high potential are light. (b)–(d) Modulus of selected characteristic wave functions, (b) at low energy in the tail of the density of states, (c) at medium energies, but still localized along equipotential lines, (d) in the vicinity of the density of states maximum, where the wave function is extended along equipotential lines representing a percolation network for electrons. (Reprinted from Kramer *et al.*, 2005 with permission from Elsevier.)

the characteristic length scale of the fluctuations. In this case, cyclotron orbits drift normal to the magnetic field direction (i.e., in the plane) and normal to the electric field, i.e., along equipotential lines ($\mathbf{E} \times \mathbf{B}$-drift). Equipotential lines at the Fermi energy will form closed contour lines around maxima and minima of the potential landscape. In a classical picture, electrons on drifting cyclotron orbits can therefore be trapped on trajectories encircling potential maxima or minima in the sample. Seen from a quantum mechanical viewpoint, the electron wave functions have maxima of their probability density distribution along equipotential lines. Figure 16.18 shows calculated wave functions in a long-range fluctuating potential. At energies in the tail of the density of states, the states are localized in minima of the potential [Fig.16.18(b)]. At higher energies, (c), the states become more extended, but the probability amplitude is nonzero only along equipotential lines. It can be shown rigorously that this is always the case in the limit of sufficiently high magnetic fields and smooth potentials (see, e.g., Kramer *et al.*, 2005). In the center of a Landau level, near the maximum in the density of states, there are percolating states that can be extended throughout the whole sample (d). The saddle points in the potential play an important role for this percolation because wave functions can branch there.

Localization in the presence of electron–electron interaction
So far we have completely neglected the interaction between electrons.

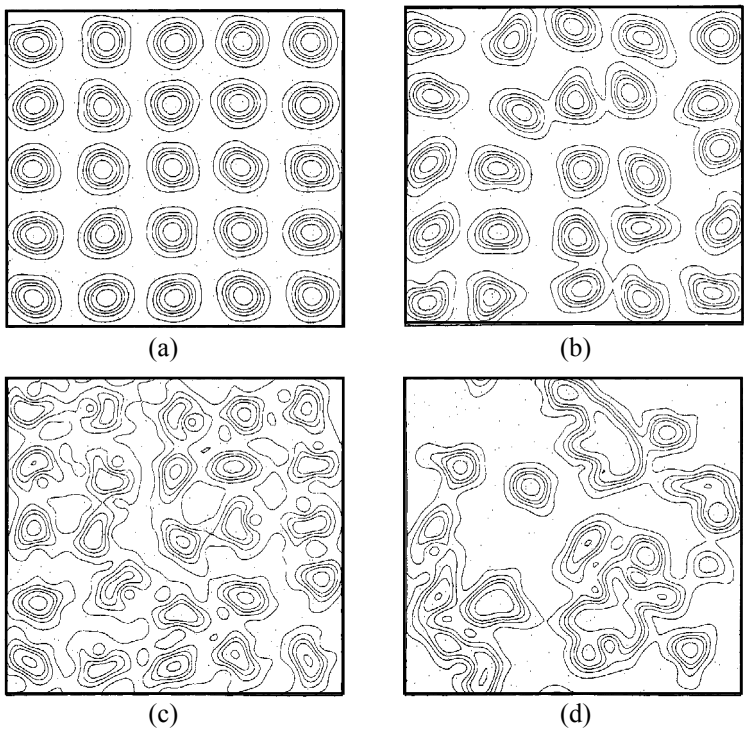


Fig. 16.19 Electron density for different strengths of the interaction potential. The interaction decreases from (a) to (d). (d) is the case of a noninteracting electron gas, (a) shows a charge density distribution reminiscent of a Wigner crystal, while (b) and (c) could be called a Wigner glass or an amorphous Wigner crystal (Aoki, 1979).

Including it leads to a rich variety of possible ground states at high magnetic fields. The interplay between spatial potential fluctuations and Coulomb interaction is an active research area even today, not only at high but also at zero magnetic field. In section 16.4 we will discuss the fractional quantum Hall effect in which interactions play a dominant role. Here we show numerical results obtained before the discovery of the quantum Hall effect. The interaction is included in the self-consistent Hartree–Fock approximation and it is assumed that the magnetic field is big enough to avoid mixing between states of neighboring Landau levels. The impurity potential is again created by randomly placed δ-scatterers in the plane. Figure 16.19 shows the total electron density for different ratios between the interaction strength and the scattering potentials. With increasing interaction strength the system evolves from an Anderson insulator via an amorphous Wigner crystal to a Wigner crystal.

16.3 The integer quantum Hall effect

The integer quantum Hall effect arises in the Hall conductivity or Hall resistivity of a two-dimensional electron gas at high magnetic field. In the previous sections we acquired the knowledge that was necessary to understand the Shubnikov–de Haas oscillations in ρ_{xx}, and we have obtained some insight into the phenomenon of vanishing ρ_{xx} and σ_{xx} at

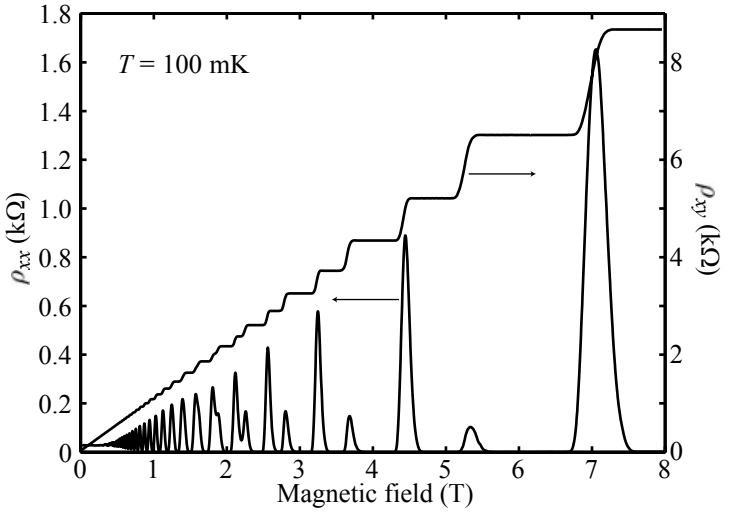


Fig. 16.20 Measurement of the longitudinal resistivity and the Hall resistivity of a two-dimensional electron gas in a GaAs/AlGaAs heterostructure as a function of the magnetic field. The measurement temperature was $T = 100\,\mathrm{mK}$.

high magnetic fields which we have attributed to the occurrence of localized states at the Fermi energy in the tails of the Landau levels. The quantum Hall effect is closely related to the zeros in ρ_{xx}, but it had not been anticipated theoretically before its discovery in 1980, almost exactly 100 years after the discovery of the classical Hall effect, by Klaus von Klitzing (von Klitzing *et al.*, 1980).

16.3.1 Phenomenology of the quantum Hall effect

If the Hall effect is measured at low temperatures (i.e., below 4.2 K) in a two-dimensional electron (or hole) gas in the quantum limit patterned into a Hall bar geometry, the Hall resistance shows a remarkable steplike increase at high fields as depicted in Fig. 16.20. The effect is found to exhibit a very high precision independent of the material in which the two-dimensional system is realized. For example, the relative difference in the plateau values between GaAs and Si systems was found to be smaller than 3.5×10^{-10}. At small magnetic fields below about 1.5 T, the Hall resistivity shows the linear increase with magnetic field expected from the classical Drude model. At higher fields, oscillatory behavior sets in, which develops into the formation of well-pronounced plateaus in the Hall resistance. The plateau values of the resistivity are with a relative accuracy of 10^{-8} given by the relation

$$\rho_{xy}^{\mathrm{plateau}} = \frac{h}{ie^2}, \tag{16.18}$$

where i is an integer number. The constant $R_{\mathrm{K}} = h/e^2 = 25\,812.807449\,\Omega$ is called the *von Klitzing constant*, or the *resistance quantum*. The ratio between experimental plateau values at $i = 1, 3, 4, 6, 8$ and that at $i = 2$, i.e., the average ratio $\langle i\rho_{xy}^{\mathrm{plateau}}(i)/2\rho_{xy}^{\mathrm{plateau}}(2)\rangle$ was found to be integer within a relative accuracy of the order of 10^{-10}.

In the magnetic field regions where plateaus occur in ρ_{xy}, the longitudinal magnetoresistance ρ_{xx} shows well-pronounced zeros (see Fig. 16.20), i.e.,

$$\rho_{xx}^{\text{plateau}} \approx 0, \tag{16.19}$$

which is the effect attributed above to the localization of states. The two equations (16.18) and (16.19) characterize the quantum Hall effect.

It was found that the precision of the quantization does not depend on the width of the Hall bars used for a large range of bar widths W between 10 µm and 100 µm within a relative accuracy of the order of 10^{-9}. It is also independent of the device mobility within the relative accuracy of the order of 10^{-9} between $\mu = 130\,000\,\text{cm}^2/\text{Vs}$ and $1.3 \times 10^6\,\text{cm}^2/\text{Vs}$, and independent of the employed fabrication process. The effect has therefore been used internationally as a standard for electrical resistance since 1990. Owing to the relation between the resistance quantum R_K and the fine structure constant α,

$$R_\text{K} = \frac{h}{e^2} = \frac{\mu_0 c}{2\alpha},$$

it can also be seen as an extremely precise method to measure the fine structure constant, which complements the usual method of measuring the anomalous magnetic moment of the electron.

Extrapolating the linear classical Hall resistance to high magnetic fields, it crosses the quantum Hall plateaus in ρ_{xy} at fields

$$B_i = \frac{n_\text{s} h}{i|e|}.$$

This means that the plateaus in ρ_{xy} and the minima in ρ_{xx} are periodic in $1/B$ and occur whenever $\nu = n_\text{s}/n_\text{L} = i$. The integer number i counts the number of occupied Landau levels and therefore reflects the filling factor ν. Plateaus in ρ_{xy} occur whenever the filling factor ν is close to an integer number i.

If we use eqs. (16.19) and (16.18) for calculating the components of the conductivity tensor from eqs. (10.15) and (10.16) we find

$$\begin{aligned} \sigma_{xx}^{\text{plateau}} &= 0 \\ \sigma_{xy}^{\text{plateau}} &= i\frac{e^2}{h}, \end{aligned} \tag{16.20}$$

i.e., there is also a quantized plateau value in σ_{xy} and σ_{xx} vanishes there. With the Hall conductivity, the Hall conductance is also quantized according to $G_{xy} = \sigma_{xy} = ie^2/h$. The constant e^2/h is called the *conductance quantum*.

The quantum Hall effect can also be seen at constant magnetic field, when the electron density is changed, e.g., with a top gate voltage. Figure 16.21 shows the original measurement from the publication of Klaus von Klitzing. The labels $n = 0, 1, 2$ indicate the Landau level quantum number. Owing to the twofold spin degeneracy and the twofold valley degeneracy in silicon inversion layers, each Landau level hosts $4n_\text{L}$ states.

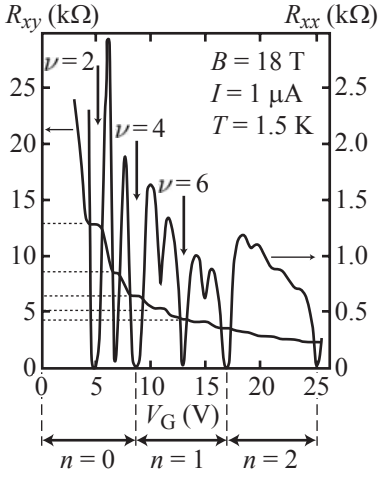


Fig. 16.21 Measurement of the quantum Hall effect in R_{xy}, and the corresponding longitudinal resistance R_{xx} as a function of the electron density which is changed via the voltage V_g on the top gate of a Si MOSFET. (Reprinted with permission from von Klitzing *et al.*, 1980. Copyright 1980 by the American Physical Society.)

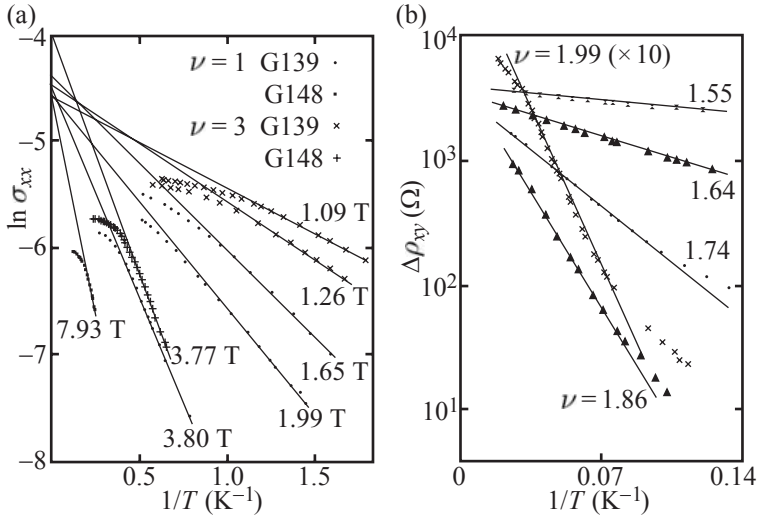


Fig. 16.22 (a) Arrhenius plots of the conductivity of two samples at the filling factors $\nu = 1$ and 3 for various densities (and therefore various magnetic fields). (Reprinted with permission from Usher *et al.*, 1990. Copyright 1990 by the American Physical Society.) (b) Arrhenius plots of $\Delta\rho_{xy} = |\rho_{xy} - h/2e^2|$ for different filling factors $\nu < 2$. (Reprinted with permission from Wei *et al.*, 1985. Copyright 1985 by the American Physical Society.)

The third R_{xx}-minimum from the left corresponds to filling factor $\nu = 4$ and therefore corresponds to one fully occupied Landau level. The spin degeneracy at the field of 18 T is lifted by the Zeeman splitting. Also the orbital degeneracy is split by a small energy gap leading to the additional minima in R_{xx}. At high density, the minima due to the small spin and valley splitting disappear.

Thermal activation of resistance minima and Hall plateaus. The width of the ρ_{xy}-plateaus and of the ρ_{xx}-minima decreases with increasing temperature. Also the resistivity values at the minima in ρ_{xx} and σ_{xx} depend on temperature. The temperature dependence of the conductance at integer filling factor $\nu = i$ turns out to exhibit activated behavior over a wide temperature range (typically from a few kelvin up to a few tens of a kelvin) following the Arrhenius law

$$\sigma_{xx} = \sigma_0 \exp\left(-\frac{\Delta_{xx}}{2k_{\mathrm{B}}T}\right), \tag{16.21}$$

where Δ_{xx} is the activation energy from the Fermi energy to the nearest unoccupied extended state near the center of the next higher Landau level (cf., Fig. 16.16). Figure 16.22(a) shows the result of an experiment on two Ga[Al]As heterostructures for the filling factors $\nu = 1$ and 3 at different electron densities. The Arrhenius plots show that the minimum conductivity is quite well described by eq. (16.21) over a certain temperature range. At the measured odd filling factors the activation energy Δ extracted from the slopes of the fits corresponds to half the spin splitting $g^\star\mu_{\mathrm{B}}B$ ($g^\star$ is the g-factor of the host material, in this case GaAs with $g^\star = -0.44$) between the highest occupied and the lowest unoccupied Landau levels. The data, however, give a much higher g-factor, namely, $|g^{\star\star}| = 7.3$. The origin of this extreme enhancement of the g-factor is the exchange interaction which favors spins to align in parallel. A stronger

Zeeman splitting increases the degree of spin polarization and therefore lowers the total energy of the system.

At very low temperatures (typically below 1 kelvin, see Fig. 16.20) where the resistivity is exponentially suppressed according to eq. (16.21), the activated description breaks down and transport arises due to hopping of electrons between localized sites in the sample, as described by other exponential laws (Briggs *et al.*, 1983).

Also, the Hall resistance ρ_{xy} shows activated behavior of the form

$$\Delta\rho_{xy}(T,B) = \left|\rho_{xy}(T,B) - \frac{h}{ie^2}\right| = \rho_0 \exp\left(-\frac{\Delta_{xy}}{2k_\mathrm{B}T}\right),$$

where i is again an integer, if the plateau value is measured at a filling factor slightly away from the integer filling factor $\nu = i$. Arrhenius plots of $\Delta\rho_{xy}(T)$ for $i = 2$ are shown in Fig. 16.22(b) as they were measured on InGaAs/InP heterostructures. With increasing difference of the filling factor from the integer value ν, the temperature dependence becomes weaker, i.e., the activation energy becomes smaller. This means that there is one point between two neighboring integer filling factors, at which the temperature dependence of $\Delta\rho_{xy}$ is close to zero.

Conditions for the observation. The quantum Hall effect is only observed in two-dimensional systems, e.g., two-dimensional electron or hole gases as they are realized in semiconductor heterostructures and quantum wells, but also in graphene layers. The effect is very pronounced if at large magnetic field the Landau level separation $\hbar\omega_\mathrm{c}$ is large compared to the Landau level broadening, i.e., if

$$\omega_\mathrm{c}\tau_q \gg 1.$$

High quality samples are therefore a crucial prerequisite. The clear observation of the effect is further possible at temperatures, where the thermal broadening $k_\mathrm{B}T$ of oscillations in the magnetotransport coefficients is smaller than the Landau gap $\hbar\omega_\mathrm{c}$. This means

$$\hbar\omega_\mathrm{c} > k_\mathrm{B}T.$$

16.3.2 Bulk models for the quantum Hall effect

The discovery of the quantum Hall effect has triggered a large variety of theoretical approaches attempting to reveal the physics behind the effect and to unravel the microscopic details of the quantum Hall state. Models considering noninteracting electrons in a perpendicular magnetic field turn out to be a good starting point. We have already seen that, at high magnetic fields, Landau levels are broadened and localized states exist in the tails of the Landau level density of states, whereas extended states exist near the density of states maximum (see Fig.16.16). At magnetic fields (or electron densities) where the Fermi energy lies in regions of localized states (i.e., near integer filling factors ν), the longitudinal

conductivity σ_{xx} [eq. (16.20)] goes to zero, as observed, for example, in Corbino geometries.

The quantization of the Hall conductance (resistance) occurs under the same conditions as the zeros in σ_{xx} and ρ_{xx}. Naively one could argue that localized electrons do not contribute to transport and therefore do not appear in the Hall resistance. Then, since $\omega_c\tau \gg 1$, $\sigma_{xy} = n_{\mathrm{mobile}}|e|/B$, where n_{mobile} is the density of nonlocalized charge carriers. This reasoning is, however, not in agreement with the experiment. Aoki and Ando discussed the influence of localization on the Hall effect (Aoki and Ando, 1981) and found that σ_{xy} does not change if the Fermi energy is in a region of localized states. They could establish the constant value of $\sigma_{xy} = ie^2/h$ only for the case in which neighboring Landau levels do not overlap, i.e., at very high magnetic fields. In this case, i turns out to be the number of Landau levels with their extended states below the Fermi energy. The physical argument for their finding is as follows: A localized state does indeed not contribute to the Hall conductance. However, extended states are influenced by the localization because all the states within a Landau level have to be orthogonal to each other. As a consequence, the extended states will have an enhanced drift velocity which exactly compensates the reduced number of mobile charge carriers. This argument is valid for systems that are infinitely extended in a plane. One therefore talks about 'bulk models' for the quantum Hall effect. The boundaries of a realistic Hall bar sample do not play a role in this description.

16.3.3 Models considering the sample edges

Real Hall bars have a finite size, and the nature of states at the edges of a sample has a profound influence on the quantum Hall effect. This was recognized by Halperin shortly after the discovery of the effect (Halperin, 1982). In order to describe states at the sample edge at high magnetic fields we introduce a confinement potential in the y-direction. Equation (16.2) is then changed to

$$\left[\frac{p_y^2}{2m^\star} + \frac{1}{2}m^\star\omega_c^2\left(y - \frac{\hbar k_x}{|e|B_z}\right)^2 + V(y)\right]\eta_{nk_x}(y) = E_n\eta_{nk_x}(y).$$

The influence of $V(y)$ on the states will depend on the detailed shape and strength of this potential. We can get some insight into the problem by assuming that it can be considered as a weak perturbation. The matrix element of this perturbation with the wave functions of the harmonic oscillator gives, at sufficiently high magnetic fields and slowly (on the length scale of the wave function extent) increasing $V(y)$, the value $\langle V(y)\rangle = V[y_0(k_x)]$. The eigenenergies are therefore given by

$$E_n(k_x) = \hbar\omega_c\left(n + \frac{1}{2}\right) + V[\hbar k_x/(|e|B_z)].$$

The potential at the sample edge lifts the Landau level degeneracy and leads to an energy dispersion $E_n(k_x)$ implying a finite group velocity of

the states given by

$$v_x = \frac{1}{\hbar}\frac{\partial V}{\partial k_x} = \left.\frac{\partial V(y)}{\partial y}\right|_{y=\hbar k_x/(eB)} \frac{1}{|e|B},$$

which is directed along equipotential lines in the x-direction. The result is identical with the $\mathbf{E} \times \mathbf{B}$ drift that has lead to $\mathbf{v}_{\text{Drift}} = \nabla V(\mathbf{r}) \times \mathbf{B}/|e|B^2$. It is also equivalent to Fig. 16.18(c) and (d) where wave functions in the interior of the sample are extended along equipotential lines.

Figure 16.23(a) shows the motion of electrons at a constant energy in a sample with a spatially inhomogeneous potential within the Hall bar and a confinement potential defining the sample edge. In the interior of the sample, the electrons are localized and encircle local extremal points of the potential. At the right edge of the sample, spatially separate channels called *edge states* are formed moving upwards. At the left sample edge, edge states move downwards. Figure 16.23(b) shows a cross-sectional view of the states along a line in the x-direction marked in (a) by two arrows. The energy of each state is plotted as a function of the center coordinate of the wave function. Figure 16.23(c) shows the same dispersion relation, but the center coordinate has now been translated into the quantum number k. Edge states relevant for the conductance are at the intersections of the Fermi energy with the dispersions of the Landau levels.

Even if the Fermi energy lies between two Landau levels in the interior of the sample where all states are localized at integer filling factor ν, extended edge states exist at the sample boundary. All edge states at a particular sample edge have electrons moving in the same direction. Electrons moving in opposite directions are spatially completely separated. As a consequence, backscattering is completely suppressed in the vicinity of integer filling factors. Voltage contacts on the same side of the Hall bar are connected via the edge states that constitute ideal dissipationless one-dimensional connections. This is the reason why the longitudinal voltage, and with it the longitudinal resistance, vanishes near integer filling factors. In contrast, contacts at opposite edges of the sample are electronically completely separated such that the Hall voltage can build up.

In contrast, at half integer filling factors, the Fermi energy coincides with the energy of percolating states in the interior of the sample, backscattering is possible, and ρ_{xx} is finite. The edge states for the highest Landau level have disappeared.

This scenario, and the presence of edge states in particular, forms the basis of the description of the quantum Hall effect in the formalism of Landauer and Büttiker to be discussed below.

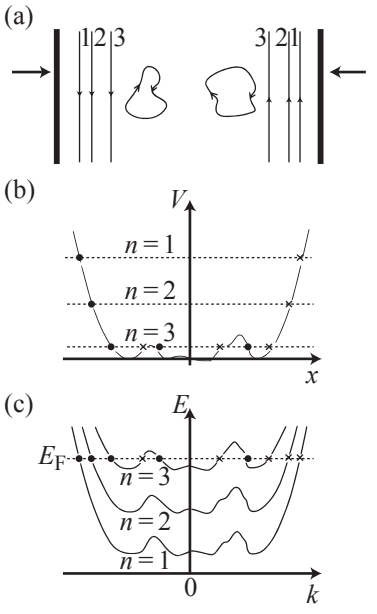


Fig. 16.23 Edge states in the quantum Hall regime in different representations. (a) Semiclassical electron motion: in the interior of the sample electrons are localized. They encircle individual maxima or minima of the potential along equipotential lines. At opposing edges of the sample, the states propagate in opposite directions (edge states). (b) Cross-sectional view along a line cut through (a) in the x-direction showing the energies of the Landau level states as a function of the orbit center. (c) Dispersion relation of the lowest three Landau levels. The edge states of the lowest Landau level are closest to the sample edge, higher Landau levels form edge states further into the bulk (Beenakker and van Houten, 1991).

16.3.4 Landauer–Büttiker picture

The quantization of the Hall conductivity in two-dimensional systems implies the quantization of the Hall conductance $G_{xy} = \sigma_{xy} = e^2/h \cdot \nu$, which has the same form as the conductance quantization in a quantum

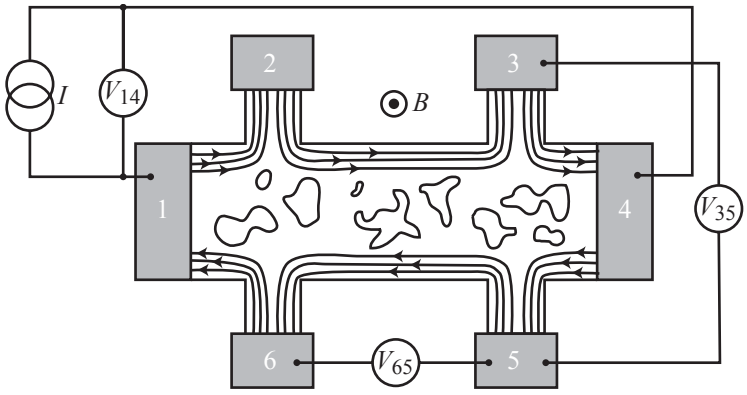


Fig. 16.24 Basic setting which allows the description of the quantum Hall effect in the framework of the Landauer–Büttiker formalism. Quantum Hall edge channels play the role of one-dimensional chiral modes connecting neighboring ohmic contacts. States in the interior of the Hall bar are localized. At integer filling factors, edge channels at opposite edges travelling in opposite directions are completely decoupled. The measurement configurations for two-terminal measurements along the entire Hall bar, and for four-terminal measurements of the longitudinal and the transverse resistance are indicated.

point contact. In section 16.3.3 we have seen that there exist extended states at the edge of the sample, even if the Fermi level lies exactly in-between two Landau levels in the bulk (integer filling factor) where all states are localized. All edge channels at a particular edge constitute a current running in the same direction. The edge channels at opposite edges are spatially well separated. This leads to suppressed coupling between edge channels propagating in opposite directions which is most effective at integer filling factors. Near half-integer filling factors states in the bulk become delocalized and tend to couple edge channels propagating in opposite directions. This situation is the basis for the description of the quantum Hall effect in the framework of the Landauer–Büttiker theory of transport which we are now going to discuss.

Figure 16.24 shows the basic picture behind the application of the Landauer–Büttiker theory to the quantum Hall effect. Close to integer filling factors ν states in the bulk of the Hall bar are localized. However, the edge states connect contacts which are neighbors in clockwise direction. The transmission of an edge state from one contact to the next in clockwise direction is perfect ($\mathcal{T} = 1$) because there is no backscattering between states travelling in opposite directions. All other transmissions are zero. The perfect transmission of edge channels is justified, because even if an electron scatters from its particular k_x-state into another k_x-state, it continues to propagate in the same direction. The reason is that the overlap and therefore the scattering matrix element is only nonzero for states with small differences in k_x (remember that k_x is related to the center coordinate of the harmonic oscillator wave function in the y-direction). For this scenario, the transmission matrix in eq. (13.10) can

be written as

$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \end{pmatrix} = \frac{e^2}{h} \begin{pmatrix} \nu & 0 & 0 & 0 & 0 & -\nu \\ -\nu & \nu & 0 & 0 & 0 & 0 \\ 0 & -\nu & \nu & 0 & 0 & 0 \\ 0 & 0 & -\nu & \nu & 0 & 0 \\ 0 & 0 & 0 & -\nu & \nu & 0 \\ 0 & 0 & 0 & 0 & -\nu & \nu \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \\ V_5 \\ V_6 \end{pmatrix}.$$

In the experiment we drive the current from contact 1 to 4, all other contacts carry no net current, because they are solely used for measurements of the voltage. This means that $I_1 = I, I_4 = -I$ and $I_2 = I_3 = I_5 = I_6 = 0$. From the above matrix equation we therefore find immediately

$$V_3 = V_2 = V_1 \text{ and } V_6 = V_5 = V_4.$$

This means that all contacts on a particular side of the Hall bar are at the same voltage. The value of this potential is determined by the current contact from which the edge channels originate. Using these relations we find for the current

$$I = \frac{e^2}{h}\nu(V_4 - V_1).$$

This relation is the complete analogue to the quantum point contact. We immediately obtain the two-terminal resistance

$$R_{2\mathrm{t}} = R_{14,14} = \frac{V_4 - V_1}{I} = \frac{h}{e^2}\frac{1}{\nu}.$$

Here $R_{ij,kl}$ denotes the resistance between contact i and j with current driven from k to l. The Hall resistance is given by

$$R_{\mathrm{H}} = R_{26,14} = \frac{V_2 - V_6}{I} = \frac{h}{e^2}\frac{1}{\nu},$$

which corresponds exactly to the observed quantized values of the Hall plateaus. The longitudinal resistance is

$$R_L = R_{65,14} = \frac{V_6 - V_5}{I} = 0,$$

again in agreement with the experiment.

We can see that our assumptions about the transmissions between the contacts based on the edge channel picture lead to the correct results for the Hall resistance and the longitudinal resistance. We can now turn the argument around and claim that the observed resistance quantization indicates that the transmissions between the contacts have exactly the values assumed in our model. However, this description does not clarify how the system realizes these values microscopically. In particular, we can ask how the current is distributed within the Hall bar on a microscopic scale. This question is still under experimental and theoretical investigation.

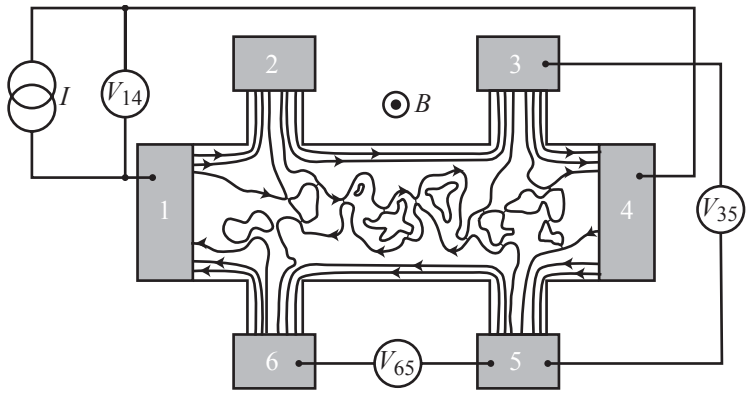


Fig. 16.25 Hall bar in a situation where the Fermi energy is at a density of states maximum. The innermost quantum Hall edge channel percolates through the sample. In this way states in the interior of the Hall bar are no longer localized. Edge channels travelling in opposite directions can couple in the bulk.

How can we, within this description, interpret the maxima in the longitudinal resistance measured as a function of density or magnetic field, and how can we understand the transitions between Hall plateaus? We can discuss the question again using the picture of edge channels. If the Fermi level coincides with the maximum in the density of states, all states at the Fermi level are extended and they percolate in the plane within the Hall bar as shown in Fig. 16.18(d). This percolation allows electrons to make their way from one edge of the bar to the other and thereby reverse their direction of motion. Figure 16.25 shows this situation schematically. It means that there is a finite probability that electrons are reflected back into the contact of their origin, and so not all elements of the tranmission matrix assumed to be zero in the above description vanish in this case. As a result the quantization of the Hall resistance disappears and the longitudinal resistance assumes a finite value.

An experiment strongly supporting the Landauer–Büttiker picture of the quantum Hall effect and the transitions between quantum Hall plateaus was made using a scanning tunneling microscopy study of a two-dimensional electron gas at high magnetic fields and a temperature of 300 mK. The two-dimensional electron gas was induced on an n-InSb(110) surface on which a hundredth of a monolayer of Cs atoms was deposited. Tip-induced band bending was minimized. This particular material is most suitable for such a study, because the small effective mass of 0.014 free electron masses sets a large cyclotron energy scale, and the large g-factor of about -51 leads to a large spin-splitting. The measured quantity is the differential conductance dI/dV as a function of tip position on the surface at fixed applied tip–sample voltage and magnetic field. This quantity is a measure of the local density of states at the surface. The tip–sample voltage is used to select the energy within the lowest Landau level at which the local density of states is determined. A magnetic field $B = 12\,\mathrm{T}$ sets the cyclotron energy scale and the filling factor. Figure 16.26(h) shows a spatially averaged dI/dV curve as a function of the tip–sample voltage. The peak in the curve resembles the density of states of the lowest (spin-resolved) Landau level.

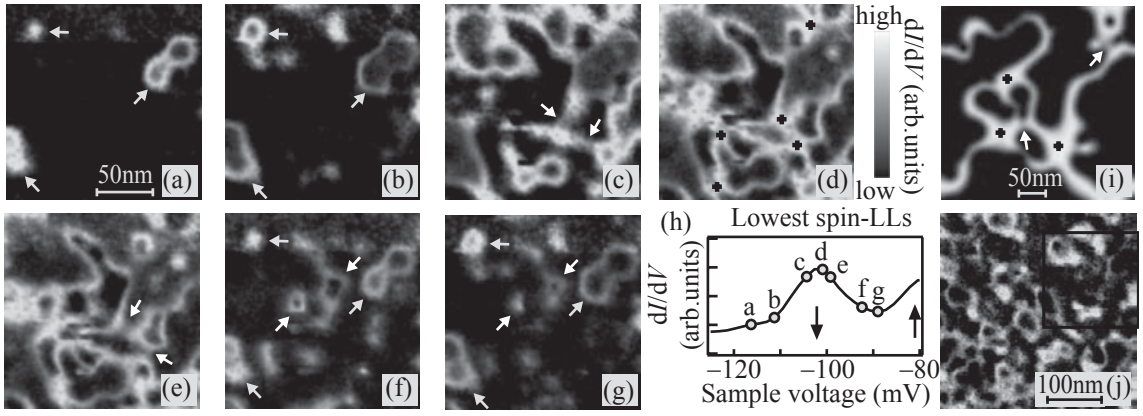


Fig. 16.26 Local density of states measured in the lowest spin-resolved Landau level of an InSb two-dimensional electron gas by scanning tunneling spectroscopy at a magnetic field $B = 12\,\mathrm{T}$. See text for details. (Reprinted with permission from Hashimoto *et al.*, 2008. Copyright 2008 by the American Physical Society.)

The local density of states measured at certain points along this curve is shown in Figs. 16.26(a)–(g). In the tails on both sides of the density of states maximum (a,g), the states tend to localize in real space along equipotential lines of the underlying disorder potential. The width of these loops is comparable with the cyclotron length $l_\mathrm{c} \approx 7.4\,\mathrm{nm}$. In regions around the density of states maximum (c–e) the states form a random network which facilitates electron transport between edges of the sample. The local density of states filaments either branch or form tunneling connections at or near saddle points of the potential. As the energy is increased from (a) to (b), loops marked by arrows increase in size, indicating that these loops encircle minima of the potential landscape. In contrast, loops in the high energy tail tend to shrink in size as the energy is increased because these states encircle maxima of the potential landscape. Figure 16.26(i) is the result of a calculation of the local density of states at the Landau level center at the same field and doping density as in the experiment, showing qualitative agreement with the experiment. Figure 16.26(j) shows an extended state at the Landau level center measured on a larger length scale.

Example: toy model for Hall cross with backscattering. In order to study the transition between Hall plateaus in more detail we consider a simple model for a Hall cross with four contacts. The transmission matrix is a 4×4 matrix of the form (13.10) with 16 elements. The current conservation sum rule requires that all elements add to zero in each column, i.e., we have the four equations

$$N_\alpha - \mathcal{R}_\alpha = \sum_{\alpha(\neq\beta)} \mathcal{T}_{\alpha\beta}.$$

If the same voltage is applied to all four contacts, we expect that there is no current flow (thermodynamic equilibrium). This leads to the sum

rule for the rows of the matrix

$$N_\alpha - \mathcal{R}_\alpha = \sum_{\beta(\neq\alpha)} \mathcal{T}_{\alpha\beta}.$$

Because only seven of these eight sum rule equations are linearly independent, they reduce the number of independent parameters from sixteen to nine. We choose these nine parameters to be $\mathcal{T}_{21}$, $\mathcal{T}_{31}$, $\mathcal{T}_{41}$, $\mathcal{T}_{32}$, $\mathcal{T}_{42}$, $\mathcal{T}_{43}$, $\mathcal{T}_{41}$, $\mathcal{T}_{13}$, $\mathcal{T}_{14}$ and find

$$\begin{aligned}
N_1 - \mathcal{R}_1 &= \mathcal{T}_{21} + \mathcal{T}_{31} + \mathcal{T}_{41} \\
N_2 - \mathcal{R}_2 &= \mathcal{T}_{32} + \mathcal{T}_{42} + \mathcal{T}_{21} - A_{13} - A_{14} \\
N_3 - \mathcal{R}_3 &= \mathcal{T}_{31} + \mathcal{T}_{32} + \mathcal{T}_{43} - A_{14} - A_{24} \\
N_4 - \mathcal{R}_4 &= \mathcal{T}_{41} + \mathcal{T}_{42} + \mathcal{T}_{43} \\
-\mathcal{T}_{12} &= -\mathcal{T}_{21} + A_{13} + A_{14} \\
-\mathcal{T}_{23} &= -\mathcal{T}_{32} + A_{13} + A_{14} + A_{24} \\
-\mathcal{T}_{34} &= -\mathcal{T}_{43} + A_{14} + A_{24},
\end{aligned}$$

where we have defined $A_{13} := \mathcal{T}_{13} - \mathcal{T}_{31}$, $A_{14} := \mathcal{T}_{14} - \mathcal{T}_{41}$, and $A_{24} := \mathcal{T}_{24} - \mathcal{T}_{42}$. Up to this point our description is general and no approximations have been made beyond those implicit in the Landauer–Büttiker theory.

In order to keep our model simple, we now approximate $\mathcal{T}_{21} = \mathcal{T}_{32} = \mathcal{T}_{43} = \mathcal{T}_{14} \equiv \mathcal{T}_{\rm cc}$ ('counterclockwise transmission'), $\mathcal{T}_{41} = \mathcal{T}_{34} = \mathcal{T}_{23} = \mathcal{T}_{12} \equiv \mathcal{T}_{\rm c}$ ('clockwise transmission'), and $\mathcal{T}_{13} = \mathcal{T}_{24} = \mathcal{T}_{31} = \mathcal{T}_{42} \equiv \mathcal{T}_{\rm s}$ ('straight transmission'). We then obtain the transmission matrix

$$\begin{pmatrix}
\mathcal{T}_{\rm cc} + \mathcal{T}_{\rm s} + \mathcal{T}_{\rm c} & -\mathcal{T}_{\rm c} & -\mathcal{T}_{\rm s} & -\mathcal{T}_{\rm cc} \\
-\mathcal{T}_{\rm cc} & \mathcal{T}_{\rm cc} + \mathcal{T}_{\rm s} + \mathcal{T}_{\rm c} & -\mathcal{T}_{\rm c} & -\mathcal{T}_{\rm s} \\
-\mathcal{T}_{\rm s} & -\mathcal{T}_{\rm cc} & \mathcal{T}_{\rm cc} + \mathcal{T}_{\rm s} + \mathcal{T}_{\rm c} & -\mathcal{T}_{\rm c} \\
-\mathcal{T}_{\rm c} & -\mathcal{T}_{\rm s} & -\mathcal{T}_{\rm cc} & \mathcal{T}_{\rm cc} + \mathcal{T}_{\rm s} + \mathcal{T}_{\rm c}
\end{pmatrix}.$$

These approximations imply the symmetry that in terms of scattering, all contacts are equivalent. In this way, the problem reduces to the three parameters $\mathcal{T}_{\rm cc}$, $\mathcal{T}_{\rm c}$, and $\mathcal{T}_{\rm s}$. Within this model we find the Hall resistance

$$R_{xy} = \frac{\mathcal{T}_{\rm cc} - \mathcal{T}_{\rm c}}{(\mathcal{T}_{\rm c} + \mathcal{T}_{\rm s})^2 + (\mathcal{T}_{\rm cc} + \mathcal{T}_{\rm s})^2}$$

and the longitudinal resistance

$$R_{xx} = \frac{\mathcal{T}_{\rm cc} + \mathcal{T}_{\rm c} + 2\mathcal{T}_{\rm s}}{(\mathcal{T}_{\rm c} + \mathcal{T}_{\rm s})^2 + (\mathcal{T}_{\rm cc} + \mathcal{T}_{\rm s})^2}.$$

Since counterclockwise scattering $\mathcal{T}_{cc}$ corresponds to the direction of the edge channels, we set

$$\mathcal{T}_{cc}(E, B) = \sum_n g[E - \hbar\omega_{\rm c}(B)(n + 1/2)],$$

where $\omega_{\rm c} = eB/m^\star$ is the cyclotron frequency. The function $g(E)$ is a smooth step function going to zero for $E \to -\infty$ and to one for $E \to \infty$

(see Fig. 16.27). The transmission $\mathcal{T}_{cc}$ counts the number of occupied Landau levels by counting the edge states existing at the Fermi energy, similar to the quantum point contact, where the number of modes below the Fermi energy are counted.

The other two parameters describe scattering channels that are only relevant if extended bulk states exist. According to the percolation theory for electrons in a spatially fluctuating potential (e.g., by Aoki and Ando, see section 16.2) this is only the case if the Fermi energy is close to the density of states maximum of a Landau level. In our model we therefore let

$$\mathcal{T}_{\mathrm{c}}(E,B) = \mathcal{T}_{\mathrm{s}}(E,B) = \sum_n h[E - \hbar\omega_{\mathrm{c}}(n+1/2)],$$

where the function $h(E)$ has a sharp maximum at $E = 0$ and goes to zero for $|E| \to \infty$ (see Fig. 16.27). This function reflects the energetic behavior of the percolation of states from one edge to another through the bulk of the sample. $\mathcal{T}_{\mathrm{c}}$ and $\mathcal{T}_{\mathrm{s}}$ therefore probe the localization length ξ of the states in the highest occupied Landau level at the Fermi energy. If ξ is larger than the size of the Hall cross, $\mathcal{T}_{\mathrm{s}}$ and $\mathcal{T}_{\mathrm{c}}$ are nonzero, and if ξ is smaller, $\mathcal{T}_{\mathrm{s}}$ and $\mathcal{T}_{\mathrm{c}}$ vanish.

In our toy model, the two functions $g(E)$ and $h(E)$ have to obey the condition $g(E) + 2h(E) \leq 1$ in order to make sure that the diagonal elements of the transmission matrix increase monotonously. Figure 16.27 shows what these functions will typically look like.

Figure 16.28 shows R_{xy} as a function of the magnetic field as calculated with the above model. The model shows nicely that within the Landauer–Büttiker formalism, ideas of bulk theories and edge state theories can be incorporated at the same time. The concerted action of the localization of electrons in the bulk, as well as the perfect transmission of electrons at the edge of the sample leads to the quantum Hall effect. However, the Landauer–Büttiker formalism does not answer

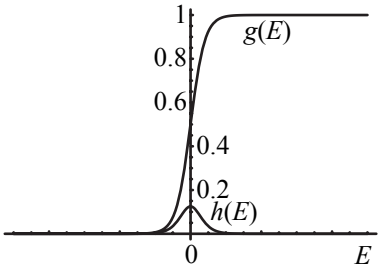


Fig. 16.27 Example for the functions $g(E)$ and $h(E)$ of the quantum Hall effect toy model.

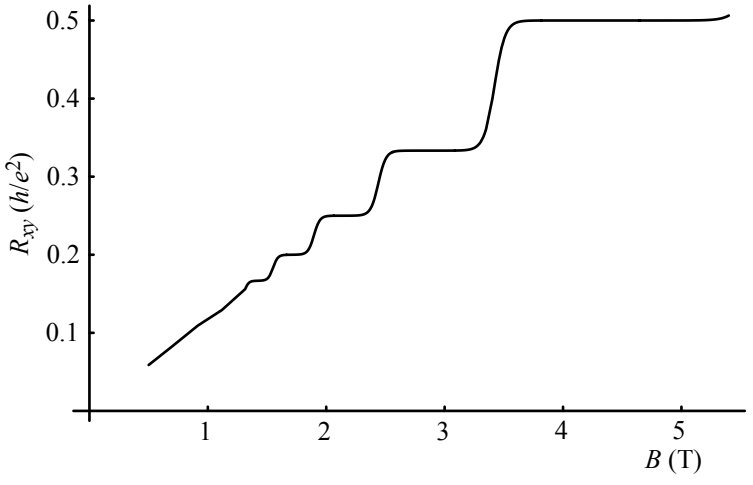


Fig. 16.28 Hall resistance calculated with the toy model described in the text. The material parameters are those of GaAs with a Fermi energy of 15 meV.

the question how the particular shape of the magnetic-field-dependent transmissions comes about microscopically, i.e., which wave functions have to be used for the description of the electrons, where exactly the currents flow in the Hall bar, and how the potential is distributed in the sample. Quantum mechanical percolation theories try to answer these questions about the microscopic origin of the quantum Hall effect.

16.3.5 Self-consistent screening in edge channels

All the above models neglected the interaction between the electrons. It has been shown that states drifting in perpendicular electric and magnetic fields either in the bulk, or at edges, are strongly affected by screening effects (Chklovskii *et al.*, 1992). The situation is schematically depicted in Fig. 16.29. The self-consistent potential does not change smoothly, but rather in step-like increments. This peculiar shape of the potential is created by the strongly oscillating local density of states. The electron density also does not change smoothly, but in step-like decrements. As a result, so-called compressible and incompressible regions are formed.

In Fig. 16.29(a) the potential at the edge of a sample is schematically depicted together with the Landau level dispersion for a noninteracting system. The local electron density near the edge behaves as depicted in (b). At positions where a Landau level that is below the Fermi energy in the interior of the sample reaches the Fermi energy, the density jumps sharply by the Landau level degeneracy $n_{\mathrm{L}} = |e|B/h$. In a quantum mechanical picture, the width of the jump is of the order of the extent of the corresponding wave function, i.e., on the scale of $l_{\mathrm{c}} = \sqrt{\hbar/|e|B}$. This behavior is strongly modified in the self-consistent treatment as shown in Fig. 16.29(c) and (d). At places where the Landau level dispersion crosses the Fermi energy, the local density of states at the Fermi energy is

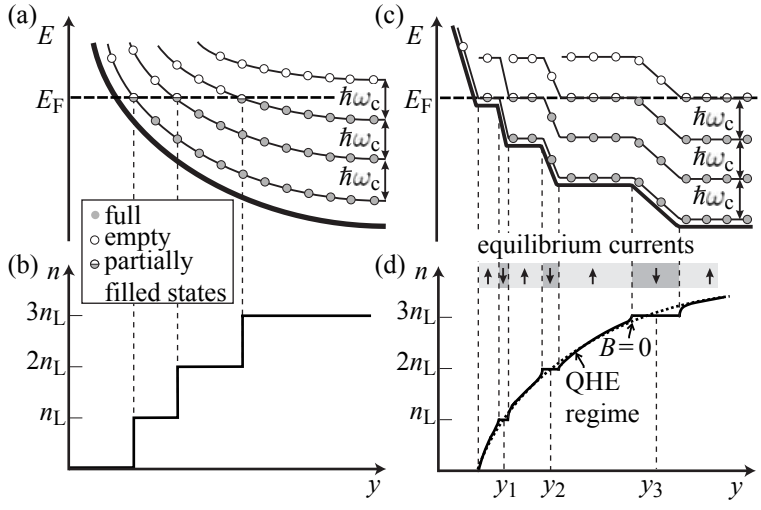


Fig. 16.29 Structure of edge states in the integer quantum Hall regime assuming spin degeneracy. (a) Adiabatic increase of Landau levels at the edge of a sample in the picture of noninteracting electrons. (b) The corresponding electron density for this case. (c) Potential and Landau level dispersion at the sample edge in the self-consistent screening model. (d) The corresponding self-consistent electron density. (Reprinted with permission from Chklovskii *et al.*, 1992. Copyright 1992 by the American Physical Society.)

high and the potential gradient can be well screened. As a consequence, the potential increase is flattened out and the electron density changes by n_L slowly and steadily over a quite large distance. These stripes are called *compressible*, because the electrochemical potential can change continuously when the electron density is increased (the compressibility κ is given by $\kappa^{-1} = n^2 \partial \mu_\mathrm{elch}/\partial n$). Between neighboring compressible stripes, the local electron density is constant, because the number of occupied Landau levels does not change. The self-consistent potential changes by about $\hbar\omega_\mathrm{c}$. These stripes between the compressible stripes are called *incompressible* because the electrochemical potential would jump here upon an increase of the electron density. Figure 16.29(d) shows that in the incompressible regions the electron density is constant.

Following Lier and Gerhardts, 1994, the separation of the incompressible stripe with integer local filling factor ν_loc from the edge is given by

$$y_{\nu_\mathrm{loc}} = \frac{d_0}{1 - \left(\frac{\nu_\mathrm{loc}}{\nu_\mathrm{bulk}}\right)^2},$$

where ν_bulk is the filling factor in the interior of the sample. The length scale d_0 measures the width of the depletion region at the sample edge. It depends on the sample fabrication and on the electron density n_s in the two-dimensional electron gas. With increasing magnetic field, the incompressible stripes move away from the sample edge until they meet in the center of the Hall bar with the corresponding stripe from the opposite sample edge and are eventually depleted. The width of incompressible stripes is given by

$$a_{\nu_\mathrm{loc}} = \frac{4 y_{\nu_\mathrm{loc}}}{\nu} \sqrt{\frac{\nu_\mathrm{loc} a_\mathrm{B}^\star}{\pi d_0}}.$$

The width of the incompressible regions increases with increasing magnetic field. Stripes that are further away from the sample edge (larger ν_loc) have a larger width than those close to the edge. Typical length scales are of the order of 100 nm.

This model of self-consistent edge channels describes an equilibrium property of the electronic system. If only small voltages are applied, this picture should not change significantly. The distribution of the equilibrium currents in the self-consistent edge channel picture were calculated in Geller and Vignale, 1995. They found

$$j_x(y) = \frac{\hbar}{2m^\star} \left[(2[\nu] + 1) + \frac{1}{\mu_\mathrm{B} \frac{\partial \mu_\mathrm{ch}^\mathrm{xc}(y)}{\partial y}} \right] \frac{\partial n(y)}{\partial y} + \frac{n(y)}{m^\star \omega_\mathrm{c}} \frac{\partial U_\mathrm{H}^\mathrm{xc}(y)}{\partial y}.$$

Here, $\mu_\mathrm{ch}^\mathrm{xc}(y)$ is an exchange energy contribution to the chemical potential, and U_H^xc is the self-consistent potential containing Hartree- and exchange interaction effects. The first term in the equation is proportional to the gradient of the electron density and therefore describes the current density in the region of a compressible stripe. The second term

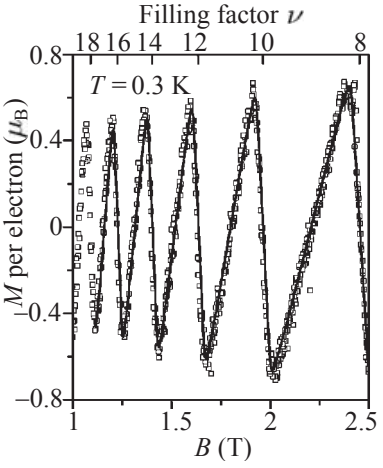


Fig. 16.30 De Haas–van Alphen oscillations of the magnetization of a two-dimensional electron gas in a Ga[Al]As sample. (Reprinted with permission from Schwarz *et al.*, 2002. Copyright 2002 by the American Physical Society.)

is proportional to the gradient of the local potential and therefore describes the current density in the incompressible regions. Because these two terms differ in sign, and the prefactors are positive, the local currents in compressible and incompressible stripes flow in opposite directions, as indicated schematically in Fig. 16.29(d).

The current in the compressible stripes leads to a diamagnetic effect, because the current that circulates at the sample edge creates a magnetic field counteracting the external field in the interior of the sample. In contrast, the current in the incompressible stripes leads to a paramagnetic effect. As compressible and incompressible regions are disappearing successively with increasing magnetic field, the whole magnetic moment of the electron gas oscillates periodic in $1/B$ around zero (Bremme *et al.*, 1999). This is the *de Haas–van Alphen effect*, a $1/B$-periodic oscillation of the magnetization of a two-dimensional electron gas in a high magnetic field. Figure 16.30 shows the result of the corresponding measurement of the magnetization.

16.3.6 Quantum Hall effect in graphene

At the end of our discussion of the integer quantum Hall effect we take a little detour and discuss the quantum Hall effect in graphene. The difference between this material and conventional semiconductor materials is the gapless linear dispersion relation near the K and K′ points of the first Brillouin zone which replaces the parabolic dispersion relation that we have discussed so far. Correspondingly, the wave functions near the K and K′ points are described by a two-component vector, resembling the two basis atoms in the primitive cell. The details of this description were described on pages 23 and 40. Graphene possesses a two-fold spin and a two-fold valley degeneracy. According to our previous discussion, quantum Hall plateaus would be expected at $\sigma_{xy} = 4ie^2/h$ with i being an integer number. The first measurements of the quantum Hall effect in graphene were reported in Novoselov *et al.*, 2005 and Zhang *et al.*, 2005. Figure 16.31 shows the result of a measurement on a Hall bar structure. Surprisingly, the plateaus do not appear at the expected integer multiples of $4e^2/h$, but at half integer values.

In order to understand this behavior, we have to investigate the Landau level structure of the graphene material (see also Ando, 2005). To this end we start from eq. (3.26) and make the transition to the analogue of the effective mass equations derived for parabolic bands in chapter 4. In a magnetic field perpendicular to the plane of the sheet we replace $\mathbf{q} = -i\nabla$ by $-i\nabla + (e/\hbar)\mathbf{A}$. We choose for the vector potential

$$\mathbf{A} = (-yB, 0, 0).$$

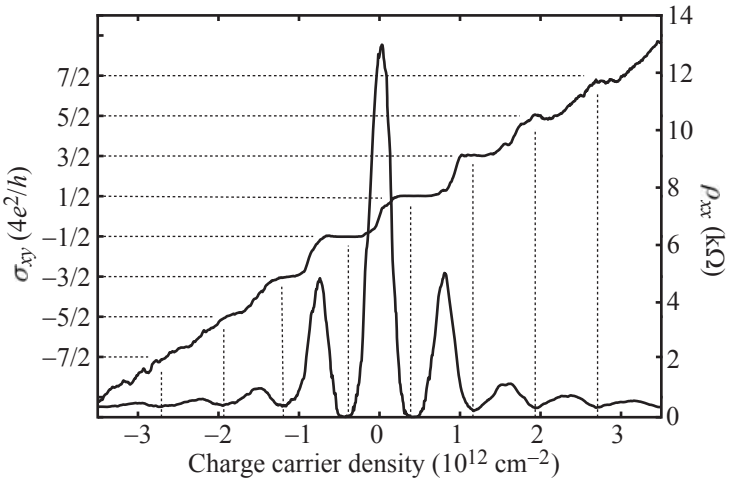


Fig. 16.31 Quantum Hall effect and longitudinal resistance measured on a graphene Hall bar at a temperature of 1.7 K.

The envelope functions are then described by the equation

$$\hbar c \begin{pmatrix} 0 & -i\partial_x - y/l_c^2 - \partial_y \\ -i\partial_x - y/l_c^2 + \partial_y & 0 \end{pmatrix} \begin{pmatrix} A_{\mathbf{q}}(\mathbf{r}) \\ B_{\mathbf{q}}(\mathbf{r}) \end{pmatrix} = E \begin{pmatrix} A_{\mathbf{q}}(\mathbf{r}) \\ B_{\mathbf{q}}(\mathbf{r}) \end{pmatrix}.$$

Letting

$$\begin{pmatrix} A_{\mathbf{q}}(\mathbf{r}) \\ B_{\mathbf{q}}(\mathbf{r}) \end{pmatrix} = e^{iq_x x} \begin{pmatrix} A_{\mathbf{q}}(y) \\ B_{\mathbf{q}}(y) \end{pmatrix}$$

and $\epsilon = E/\hbar c$ we obtain

$$\begin{pmatrix} 0 & q_x - y/l_c^2 - \partial_y \\ q_x - y/l_c^2 + \partial_y & 0 \end{pmatrix} \begin{pmatrix} A_{\mathbf{q}}(y) \\ B_{\mathbf{q}}(y) \end{pmatrix} = \epsilon \begin{pmatrix} A_{\mathbf{q}}(y) \\ B_{\mathbf{q}}(y) \end{pmatrix}.$$

We decouple this system of equations by multiplying both sides of one equation by ϵ and inserting it into the other and vice versa. This leads from

$$\begin{aligned} (q_x - y/l_c^2 - \partial_y) B_{\mathbf{q}}(y) &= \epsilon A_{\mathbf{q}}(y) \\ (q_x - y/l_c^2 + \partial_y) A_{\mathbf{q}}(y) &= \epsilon B_{\mathbf{q}}(y) \end{aligned}$$

to

$$\begin{aligned} (q_x - y/l_c^2 - \partial_y)(q_x - y/l_c^2 + \partial_y) A_{\mathbf{q}}(y) &= \epsilon^2 A_{\mathbf{q}}(y) \\ (q_x - y/l_c^2 + \partial_y)(q_x - y/l_c^2 - \partial_y) B_{\mathbf{q}}(y) &= \epsilon^2 B_{\mathbf{q}}(y). \end{aligned}$$

Further simplifications give

$$\begin{aligned} \left[-\partial_y^2 + (q_x - y/l_c^2)^2\right] A_{\mathbf{q}}(y) &= (\epsilon^2 - 1/l_c^2) A_{\mathbf{q}}(y) \\ \left[-\partial_y^2 + (q_x - y/l_c^2)^2\right] B_{\mathbf{q}}(y) &= (\epsilon^2 + 1/l_c^2) B_{\mathbf{q}}(y). \end{aligned}$$

These are harmonic oscillator equations with solutions for the energies

$$E = \pm\sqrt{2e\hbar c^2 B(n_A + 1)} \quad \text{for } n_A = 0, 1, 2, \ldots$$
$$E = \pm\sqrt{2e\hbar c^2 B n_B} \quad \text{for } n_B = 0, 1, 2, \ldots$$

It is interesting to note that the lowest Landau level occurs at energy $E = 0$, but only solutions for one pseudospin exist, i.e., the other component is zero. All higher Landau levels are composed of states with both pseudospins mixed. The states of the Landau levels are the usual harmonic oscillator wave functions displaced by $q_x l_c^2$ in the plane. Note, however, that two pseudospin wave functions occurring at the same energy have different Landau level quantum numbers. The Landau level degeneracy is, as in any other two-dimensional system, given by $n_\mathrm{L} = eB/h$. It is straightforward to verify that doing the same analysis at $\mathbf{K}'$ leads to the same results, but with the roles of lattice sites A and B interchanged. Each Landau level state can therefore be occupied with four electrons, i.e., two opposite spins and two opposite valleys. Therefore all Landau levels at the same energy can take $4n_\mathrm{L}$ electrons.

The Landau level at zero energy is a special property of the graphene band structure. It leads to the resistance maximum at zero charge carrier density in Fig. 16.31. The half integer filling factors at the plateaus can be understood from the following argument: If we consider the Landau-levels to be symmetrically broadened by disorder, half of its states belong to the valence band, and half to the conduction band. Starting from the charge neutrality point we will have the first plateau, when the zero energy Landau level is completely filled, i.e., when we have $2n_\mathrm{L}$ electrons in the conduction band (rather than $4n_\mathrm{L}$). This fact leads to the shift of the first plateau from $4e^2/h$ to $2e^2/h$, and any higher plateaus will correspondingly appear at quantized values

$$\sigma_{xy} = \frac{4e^2}{h}\left(i - \frac{1}{2}\right)$$

for integer i. For other aspects of the zero-energy Landau level involving Berry's phase we refer the reader to the review Ando, 2005.

Another interesting aspect of the quantum Hall effect in graphene is the absolute energy spacings at experimentally achievable magnetic fields. For example, at a magnetic field $B = 10\,\mathrm{T}$, the energy separation between the first and the zero energy Landau level is given by $\Delta E_{10} = 115\,\mathrm{meV}$ which is more than four times the thermal energy $k_\mathrm{B}T$ at room temperature. This huge energy scale has made the observation of the quantum Hall effect at room temperature possible (Novoselov *et al.*, 2007).

16.4 Fractional quantum Hall effect

16.4.1 Experimental observation

In 1982, soon after the discovery of the integer quantum Hall effect in a silicon MOS structure, Tsui *et al.*, 1982, found plateaus in the

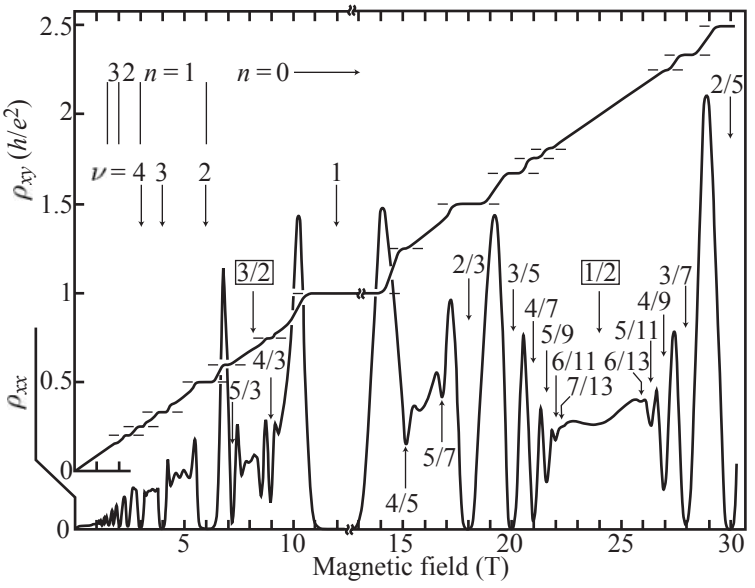


Fig. 16.32 Measurement of the fractional quantum Hall effect in a GaAs/AlGaAs heterostructure at the temperature $T = 100\,\mathrm{mK}$. The numbers in the figure indicate the filling factors. (Reprinted with permission from Willett *et al.*, 1987. Copyright 1987 by the American Physical Society.)

Hall resistance beyond the integer quantum Hall effect that occurred at fractional filling factors.

Figure 16.32 shows the result of a measurement of the fractional quantum Hall effect. The phenomenon can only be observed in samples with very high mobility [typically more than $10^6\,\mathrm{cm}^2/\mathrm{Vs}$]. The theory of the fractional quantum Hall effect was strongly influenced by the ideas of Laughlin (Laughlin, 1983), who considered interactions between electrons to be crucial for the occurrence of the effect. Tsui, Störmer, and Laughlin were awarded the Nobel prize for physics in 1998 for the experimental discovery and its theoretical description.

Phenomenology of the fractional quantum Hall effect. Phenomenologically, the fractional quantum Hall effect is very similar to the integer effect. The plateaus in the Hall resistance occur at values

$$\rho_{xy}^{\mathrm{plateau}} = \frac{h}{e^2} \cdot \frac{1}{p/q},$$

where q and p are both integers, but q is bound to be odd in (almost) all cases (see below for so-called even denominator fractions). At the same magnetic fields where plateaus exist in ρ_{xy}, the longitudinal resistivity ρ_{xy} shows minima that approach zero at sufficiently low temperatures (see $\nu = p/q = 2/5$ or $2/3$ in Fig. 16.32). The resistivity values of ρ_{xx} in the minima were found to exhibit an exponential temperature dependence, i.e., $\rho_{xx}(T) \propto \exp(-\Delta_{p/q}/k_{\mathrm{B}}T)$ as in the integer quantum Hall effect, indicating the formation of electronic ground states separated from the lowest excitations by an energy gap $\Delta_{p/q} \ll \hbar\omega_{\mathrm{c}}$. It is commonly believed that this energy gap is a result of electronic correlations

brought about by the Coulomb interaction leading to a rich substructure within each Landau level.

The fractional quantum Hall effect can only be observed if the scale of the disorder potential in the two-dimensional electron gas, and the temperature, are smaller than the energy scale $\Delta_{p/q}$. Nevertheless, small disorder is generally believed to be crucial for the observation, because it serves to localize quasiparticles and brings about finite-width Hall plateaus similar to the integer quantum Hall effect.

16.4.2 Laughlin's theory

In the range of filling factors $\nu < 1$ all electrons are in the lowest spin-polarized Landau level, i.e., all spins are oriented in parallel. In this high magnetic field range, no other Landau levels play a role, and the kinetic energy is the same for all electrons and therefore an irrelevant constant. As a consequence, the Coulomb interaction between electrons becomes dominant for the dynamics of the electrons. Bob Laughlin published a theory of the fractional quantum Hall effect in 1983 (Laughlin, 1983) in which he suggested the many-body ground state wave function for filling factor $\nu = 1/m$ ($m > 0$ is an odd number)

$$\psi_{1/m} = \prod_{j<k} (z_j - z_k)^m e^{-\sum_l |z_l|^2/4}. \tag{16.22}$$

Here, $z_j = x_j + iy_j$ is the position of the jth electron in complex notation. The exponential factors in this wave function correspond to the ground state wave function of noninteracting electrons in a magnetic field. The prefactor (Laughlin–Jastrow factor) creates nodes in the wave function for the case that two electron positions coincide. This incorporates spatial correlations between the electrons minimizing their Coulomb interaction. The number m has to be odd, in order to ensure that the wave function changes sign when two particles are interchanged. We can visualize how this wave function works in principle by keeping all electron coordinates z_j in eq. (16.22) with $j > 0$ fixed and plotting the resulting wave function as a function of z_0. Figure 16.33 shows an example for filling factor $\nu = 1/3$. In (a) the squared modulus of the resulting wave function is plotted on a logarithmic color scale (printed here in grayscale). The probability density for an electron vanishes at those locations where the other electrons sit. In (b) the phase of the wave function is shown. If the free electron travels on a closed path around one of the fixed electrons, it changes its phase by $2\pi/\nu$, i.e., by 6π in our case of $\nu = 1/3$. This is equivalent to the Aharonov–Bohm phase acquired if an electron encircles three magnetic flux quanta. We therefore talk about composite particles where three flux quanta are attached to each electron. The nodes in the wave function corresponding to the singularity points of the phase are called *vortices*. An important consequence of Laughlin's theory is the insight that quasiparticle excitations of the system at filling factor $\nu = 1/m$ carry the fractional charge $q_{eff} = -|e|/m$. In this picture, the quantization of G_{xy} occurs

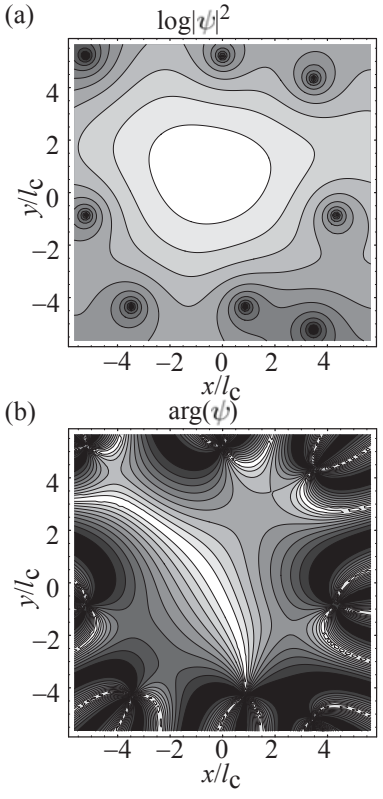


Fig. 16.33 (a) Conditional probability density for an electron with coordinates $z_0 = (x+iy)/l_c$, if the coordinates of all other electrons are fixed. (b) The phase of the corresponding wave function.

at integer filling factors of the conductance quantum $-|e|q_{eff}/h$. Much as for the integer quantum Hall effect, together with $G_{xx} \approx 0$ we obtain the quantized Hall resistivities $h/(eq_{eff}) \cdot 1/\nu_{eff}$.

16.4.3 New quasiparticles: composite fermions

The understanding of the analogy between the fractional and the integer quantum Hall effect has made big progress with the introduction of new quasiparticles, the so-called *composite fermions* (overviews are, for example, found in Jain 2000; Heinonen 1998).

Phenomenologically we observe a striking similarity of the ρ_{xx}-oscillations left and right of $\nu = 1/2$ with Shubnikov–de Haas oscillations around zero magnetic field. The same similarity is found for the Hall resistance ρ_{xy} around $\nu = 1/2$ and around zero magnetic field. This similarity has motivated the idea of describing the fractional quantum Hall effect around filling factor $\nu = 1/2$ as the integer quantum Hall effect of new quasiparticles for which the effective magnetic field B_{eff} vanishes at $\nu = 1/2$.

In this picture the magnetic field at filling factor $\nu = 1/2$ is eliminated from the description, by attaching two magnetic flux quanta $\phi_0 = h/|e|$ to each electron (Jain, 1989). This is based on the fact that at filling factor $\nu = 1/2$ the number of flux quanta per unit area is exactly twice the number of electrons per unit area, i.e.,

$$n = \frac{B_{1/2}}{2h/|e|}.$$

The new quasiparticle, called a composite fermion, incorporates the external magnetic field and eliminates it from the description of the quasiparticles at filling factor 1/2.

At magnetic fields $B \neq B_{1/2}$, the composite fermions experience the effective field

$$B_{\text{eff}} = B - B_{1/2}.$$

Landau levels for composite fermions result with a Landau level splitting (energy gap) given by $\hbar\omega_c^{\text{eff}}$ and the degeneracy factor eB_{eff}/h. Correspondingly there is an integer quantum Hall effect for composite fermions at integer effective filling factors

$$\nu_{\text{eff}} = \frac{n}{|e|B_{\text{eff}}/h}.$$

From this equation we find the relation between ν_{eff} and the filling factor ν for electron Landau levels:

$$\nu = \frac{\nu_{\text{eff}}}{1 + 2\nu_{\text{eff}}}.$$

Inserting integer values for ν_{eff} we obtain the fractional values of ν as shown in Table 16.1. The plateaus in ρ_{xy} and minima in ρ_{xx} at these fractional values of filling factors can therefore be interpreted as the integer quantum Hall effect of composite fermions at the filling factor

Table 16.1 Correspondence between fractional filling factors ν for electrons and integer filling factors ν_{eff} for composite fermions. Negative filling factors correspond to negative magnetic field values.

ν_{eff}	ν
1	1/3
−1	1
2	2/5
−2	2/3
3	3/7
−3	3/5

$\nu_{\rm eff}$. The series in Table 16.1 can be arbitrarily continued. In particular, all fractional filling factors have integer numerators and odd integer denominators.

The ground state wave function at filling factor ν is in this picture given by

$$\psi_\nu = \prod_{j<k}(z_j - z_k)^2 \phi_{\nu_{\rm eff}}. \tag{16.23}$$

Here, $z_j = x_j + iy_j$ is the position of the jth electron in complex notation and the $\phi_{\nu_{\rm eff}}$ are Slater determinants of single-particle wave functions of noninteracting electrons at filling factor $\nu_{\rm eff}$. In the wave function ψ_ν the factor $\prod_{j<k}(z_j - z_k)^2$ makes sure that each electron sees a node in the wave function at the position of each other electron. This implies that the electrons avoid each other and the probability of finding two electrons at the same place is zero. In this way the Coulomb energy is minimized. The Slater determinant makes sure that the wave function is antisymmetric under the exchange of two electrons.

If the jth electron encircles the kth, an additional phase $2 \times 2\pi$ is created. Therefore, a node in the wave function is equivalent to two flux quanta (cf., the discussion of Laughlin's wave function). How does it happen that the equation of motion for composite fermions does not contain the external magnetic field B, but the effective field $B_{\rm eff}$? In order to answer this question we consider a closed path enclosing the area A. If we move an electron counterclockwise along the path it accumulates the Aharonov–Bohm phase $\Delta\varphi_{\rm AB} = -2\pi BA/(h/|e|)$. At the same time, the number of nodes within the area A is equal to nA leading to a phase $\Delta\varphi_{\rm nodes} = 2\times 2\pi nA$. The total phase $\Delta\varphi = -2\pi BA/(h/|e|) + 4\pi nA$ can be interpreted as the Aharonov–Bohm phase in the effective magnetic field $B_{\rm eff} = B - B_{1/2}$. This means that $B_{\rm eff}$ is the relevant magnetic field for the motion of the composite fermion. Instead of talking about nodes of the wave function we can visualize composite fermions as electrons with two attached flux quanta.

At the filling factor $\nu = 1/3$ ($\nu_{\rm eff} = 1$) the wave function (16.23) is

$$\begin{aligned}
\psi_{1/3} &= \prod_{j<k}(z_j - z_k)^2 \phi_{\nu_{\rm eff}=1} \\
&= \prod_{j<k}(z_j - z_k)^2 \left[\prod_{j<k}(z_j - z_k) e^{-\sum_l |z_l|^2/4}\right] \\
&= \prod_{j<k}(z_j - z_k)^3 e^{-\sum_l |z_l|^2/4},
\end{aligned}$$

i.e., it reproduces Laughlin's wave function (16.22). In the picture of composite fermions it is interpreted as the wave function for a completely filled composite fermion Landau level.

In lowest order, interactions between composite fermions are neglected and the picture of noninteracting quasiparticles in the effective magnetic field $B_{\rm eff}$ is used. At $B_{\rm eff} = 0$, i.e. at $\nu = 1/2$ ($\nu_{\rm eff} = \infty$) the composite fermions form a Fermi sea up to the Fermi energy $E_{\rm F}^{\rm eff} = 2E_{\rm F}$. The

factor 2 results from the lifted spin degeneracy at $\nu > 1$. At $B_{\rm eff} \neq 0$ composite fermion Landau levels form which lead to plateaus in the Hall resistance and minima in the longitudinal resistance.

A comparison of exact diagonalization calculations of the problem with up to 12 electrons, and the ground state energies of the composite fermion model at various filling factors, show that the wave function (16.23) reproduces the exact ground state energy typically within 0.1%. This shows that this wave function is an excellent approximation for the problem.

The wave function (16.23) describes the correlated electronic system also very well at magnetic fields at which no fractional quantum Hall state exists. In particular this is true for magnetic fields around $B_{1/2}$. For example, magnetic focusing experiments have demonstrated that composite fermions have a semiclassical cyclotron radius

$$R_{\rm c}^{\rm eff} = \frac{\hbar k_{\rm F}^{\rm eff}}{|e|B_{\rm eff}}.$$

As a result of the spin polarization, $k_{\rm F}^{\rm eff} = \sqrt{4\pi n}$. In addition, the experiment shows that composite fermions carry the charge $-|e|$. As a result, the theory gives a coherent description of the dynamics of electrons between $\nu = 1$ and $\nu = 1/3$ in the picture of noninteracting composite fermions.

So far we have assumed that all spins at high magnetic fields are aligned in parallel. There are experiments at smaller fields showing that composite fermions carry a spin 1/2 like electrons. The Zeeman splitting between composite fermion Landau levels of opposite spin orientation has been measured in a tilted magnetic field (Melinte *et al.*, 1999; Kukushkin *et al.*, 1999).

An effective mass of composite fermions arises purely from interaction effects, because the kinetic energy of the electrons is irrelevant. It is therefore not at all related to the effective mass of electrons at the Γ-minimum or with the free electron mass. The composite fermion mass was measured, for example, in cyclotron resonance experiments (Kukushkin *et al.*, 2002). The effective mass absorbs the main part of the Coulomb interaction between the electrons. It can, however, be shown that a small interaction remains between composite fermions.

16.4.4 Composite fermions in higher Landau levels

Above we have discussed composite fermions in the lowest spin-polarized Landau level. Figure 16.34 shows a measurement of the longitudinal resistivity of a very high quality ($\mu = 12 \times 10^6\,{\rm cm}^2/{\rm Vs}$) Ga[Al]As heterostructure in the range of filling factors around $\nu = 3/2$, where the Fermi energy lies in the second Landau level. A sequence of fractional filling factor minima are observed at $\nu = 4/3$, 7/5, 10/7, 13/9, 14/9, 11/7, 8/5, and 5/3, resembling the formation of ground states with energy gaps.

It turns out that the composite fermion description remains valid at

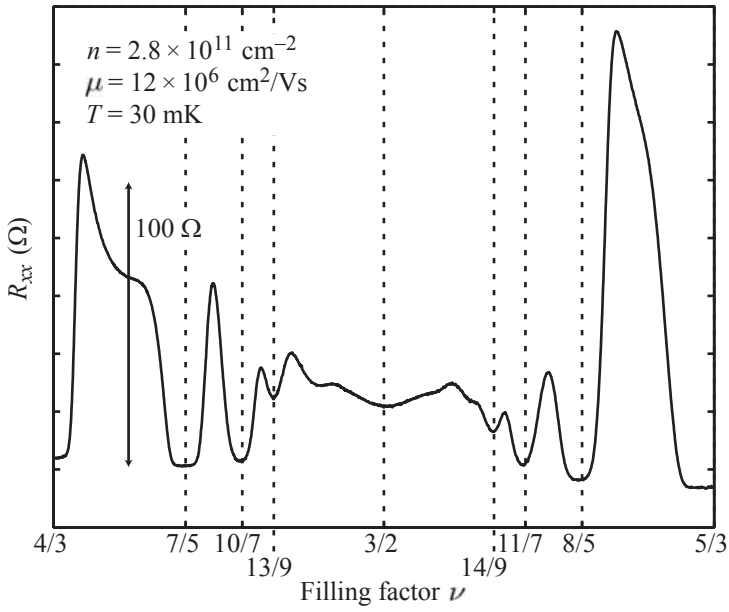


Fig. 16.34 Measurement of the fractional structure of the second Landau level around filling factor $\nu = 3/2$ in a GaAs/AlGaAs heterostructure at the temperature $T = 30$ mK. Measurement courtesy of Ch. Charpentier, U. Gasser, K. Ensslin, ETH Zurich; High mobility wafer: Werner Wegscheider, University of Regensburg.

filling factors $2 \geq \nu \geq 1$. For example, at filling factor $\nu = 3/2$, the lowest Landau level is completely filled, whereas the second Landau level is exactly half filled. We can now attach two flux quanta to each of the electrons in the half-filled Landau levels. These composite fermions will then move in an effective magnetic field $B_{\text{eff}} = 0$ at filling factor $\nu = 3/2$. At this filling factor, the density of composite fermions is only one third of the total electron concentration. If the magnetic field is slightly increased from $\nu = 3/2$, the composite fermion density decreases, as more and more electrons can be accommodated in the lowest Landau level. It can be shown that the effective magnetic field seen by these composite fermions is given by

$$B_{\text{eff}} = 3(B - B_{3/2}).$$

As a consequence, the relation between filling factor and effective filling factor of composite fermion Landau levels is

$$\nu_{\text{eff}} = \frac{\nu - 1}{3 - 2\nu} \Leftrightarrow \nu = \frac{3\nu_{\text{eff}} + 1}{2\nu_{\text{eff}} + 1}.$$

This leads to the fractional filling factors shown in Table 16.2. They are in agreement with the filling factors at which minima in the longitudinal resistivity are observed in Fig. 16.34.

Table 16.2 Correspondence between fractional filling factors ν for electrons and integer filling factors ν_{eff} for composite fermions in the second Landau level. Negative filling factors correspond to negative effective magnetic field values.

ν_{eff}	ν
1	4/3
−1	2
2	7/5
−2	5/3
3	10/7
−3	8/5

16.4.5 Even denominator fractional quantum Hall states

The achievement of ever increasing mobilities in two-dimensional electron gases formed in Ga[Al]As heterostructures, and of lower and lower

temperatures at which transport experiments were performed, allowed the observation of gapped ground states with smaller and smaller activation energies $\Delta_{p/q}$. In 1987 a quantized Hall plateau at the even denominator filling factor $\nu = 5/2$ was experimentally found (Willett *et al.*, 1987). Later experiments demonstrated a very precise quantization of the Hall resistance at $T = 4\,\mathrm{mK}$ to within 2 ppm, and an activation energy gap $\Delta_{5/2}$ of about 110 mK in a sample with a mobility of $17 \times 10^6\,\mathrm{cm}^2/\mathrm{Vs}$ (Pan *et al.*, 1999). The results from these measurements are shown in Fig. 16.35. The additional odd denominator fractional quantum Hall states with clear plateaus at $\nu = 7/3$, and $8/3$ correspond to states of noninteracting composite fermions with effective filling factors

$$\nu_{\mathrm{eff}} = \frac{\nu - 2}{5 - 2\nu} \Leftrightarrow \nu = \frac{5\nu_{\mathrm{eff}} + 2}{2\nu_{\mathrm{eff}} + 1},$$

living in an effective magnetic field

$$B_{\mathrm{eff}} = 5(B - B_{5/2}).$$

It has been argued that with the residual interaction between composite fermions, possibilities for novel composite fermion phases arise, such as superconductivity or a Wigner crystal. Currently, the most likely candidate for the $\nu = 5/2$ ground state is a state in which composite fermions in the third Landau level pair up similar to the formation of Cooper pairs in the BCS theory of superconductivity. In this picture, the state is seen as a BCS-like ground state of bosonic composite fermion pairs. An interesting aspect of the states at $\nu = 5/2$ is that the quasiparticle excitations above the ground state are neither bosons nor fermions, but particles known as nonabelian anyons. For this reason, the 5/2-state is discussed theoretically as a possible candidate for the realization of fault-tolerant topological quantum computation [see, e.g., Nayak *et al.*, 2008]. Experiments on this state are very demanding because of the required ultra-high-mobility samples, and the low electronic temperature.

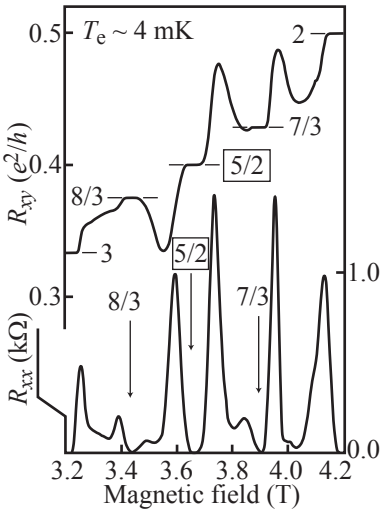


Fig. 16.35 Measurement of the fractional structure of the third Landau level around filling factor $\nu = 5/2$ in a GaAs/AlGaAs heterostructure at the temperature $T = 4\,\mathrm{mK}$. (Reprinted with permission from Pan *et al.*, 1999. Copyright 1999 by the American Physical Society.)

16.4.6 Edge channel picture

A theory of edge channels for the regime of the fractional quantum Hall effect has been developed in Beenakker, 1990 and MacDonald, 1990. It allows the application of the Landauer–Büttiker formalism to the fractional quantum Hall effect.

The basic idea is that, as the electron density tends to zero towards the sample edge, there are regions where the electron gas has local fractional filling factors of 1/3, 2/3, etc. If the potential gradient in these regions is small (i.e., the change of the potential is much smaller than the fractional energy gap over a length scale of the magnetic length), then the energy of the electron gas in these regions is lowered and the fractional energy gap can form. The gap formation is associated with the formation of a spatially extended region of constant filling factor, i.e., an incompressible stripe along the edge. Assume there is a stripe with $\nu = 1/3$ closer to

the edge, and $\nu = 1$ in the bulk of the sample. Increasing the chemical potential slightly by $\Delta\mu$ will lead to an additional electron density

$$\Delta n = \left.\frac{\delta n}{\delta \mu}\right|_{\phi=\text{const.}} \Delta\mu = -\left.\frac{\delta n}{\delta \phi}\right|_{\mu=\text{const.}} \Delta\mu$$

in the regions between constant filling factors. This additional charge moves along lines of constant potential, i.e., between the channels of constant filling factor, with drift velocity

$$v_\mathrm{d} = -\frac{1}{|e|B}\frac{\partial \phi}{\partial x}$$

and gives a current density

$$j = -|e|\Delta n v_\mathrm{d} = -|e|\left.\frac{\delta n}{\delta \phi}\right|_{\mu=\text{const.}} \Delta\mu \frac{1}{|e|B}\frac{\partial \phi}{\partial x} = -\frac{|e|}{h}\Delta\mu\frac{\partial \nu}{\partial x}.$$

The corresponding current is obtained by integrating over x to be

$$I = -\frac{|e|}{h}\Delta\mu\Delta\nu,$$

where $\Delta\nu$ is the difference between the filling factor regions enclosing the current carrying region, in our example $\Delta\nu = 1 - 1/3 = 2/3$. The conductance of such an edge current can therefore be written as

$$G = \frac{e^2}{h}\Delta\nu.$$

For edge channels in the regime of the integer quantum Hall effect we always have $\Delta\nu = 1$ and Büttiker's original edge channel picture for the integer quantum Hall effect is recovered. In our example, we get a conductance of $G = 2e^2/3h$.

16.5 The electronic Mach–Zehnder interferometer

In this chapter we have been looking at the physics of the quantum Hall effect. The formation of edge channels was one of the important concepts for its understanding. The edge states can be seen as the analogues to the one-dimensional modes in an ideal quantum point contact constriction without backscattering. The suppression of backscattering in the quantum Hall regime has been found to be very robust, because counterpropagating edge channels are separated in real space by macroscopic distances. Edge states can be seen as unidirectional electron wave guides in which electron waves propagate coherently as laser light would propagate in an optical fiber. Therefore, concepts from optics can be transferred to semiconductor nanostructure by exploiting this analogy. One particular example is the electronic Mach–Zehnder interferometer.

Interferometry in optics has brought about a number of setups for the observation of interference effects. The ring geometry can be seen as the electronic version of a double-slit experiment. The Fabry–Perot interferometer is another example which can be seen as having an electronic relative, as we will see in a later chapter on quantum dot structures. Here we will show, as another example of the realization of interference in nanostructures, the electronic version of the Mach–Zehnder interferometer. The basic principle of this interferometer is shown in Fig. 16.36. Coherent monochromatic light is sent from a source S towards a beam splitter. At the beam splitter the light has a transmission amplitude t_1 for straight transmission, and a reflection amplitude r_1 for reflection in the direction normal to the incident beam. In order to conserve the flux, these two amplitudes obey the relation $|r_1|^2 + |t_1|^2 = 1$. Both partial waves will then further propagate to a mirror where they are reflected towards the same second beam splitter, where they are made to interfere. The total transmission amplitude for a photon to be transmitted into the observation direction A is given by

$$t_{\mathrm{A}} = r_1 e^{i\varphi_\alpha} r_2 + t_1 e^{i\varphi_\beta} t_2'$$

(again we have $|r_2|^2 + |t_2|^2 = 1$ and $|r_2'|^2 + |t_2'|^2 = 1$), whereas the amplitude for transmission into direction B is

$$t_{\mathrm{B}} = r_1 e^{i\varphi_\alpha} t_2 + t_1 e^{i\varphi_\beta} r_2'.$$

The phases $\varphi_{\alpha/\beta}$ depend on the optical path length of paths α and β. We observe that in general, r_1 and t_1 are complex numbers, but their phases can be absorbed in the phases $\varphi_{\alpha/\beta}$ for simplifying further calculations. The primed transmission and reflection amplitudes occur because there are two beams from different directions incident on the second beam splitter. We will denote the phase angles of the respective amplitudes as θ_{r}, θ_{t}, θ_{r}', and θ_{t}'. The corresponding transmission probabilities are then given by

$$\begin{aligned}
\mathcal{T}_{\mathrm{A}} &= |r_1 r_2|^2 + |t_1 t_2|^2 + 2|r_1 t_1 r_2 t_2| \cos(\varphi_\alpha - \varphi_\beta + \theta_{\mathrm{r}} - \theta_{\mathrm{t}}') \\
\mathcal{T}_{\mathrm{B}} &= |r_1 t_2|^2 + |t_1 r_2|^2 + 2|r_1 t_1 t_2 r_2| \cos(\varphi_\alpha - \varphi_\beta + \theta_{\mathrm{t}} - \theta_{\mathrm{r}}').
\end{aligned}$$

Describing a single beam splitter with an S-matrix of the form (13.14), we find $\theta_{\mathrm{t}} - \theta_{\mathrm{r}}' = \theta_{\mathrm{r}} - \theta_{\mathrm{t}}' + \pi$ leading to

$$\begin{aligned}
\mathcal{T}_{\mathrm{A}} &= |r_1 r_2|^2 + |t_1 t_2|^2 + 2|r_1 t_1 r_2 t_2| \cos(\varphi_\alpha - \varphi_\beta + \theta_{\mathrm{r}} - \theta_{\mathrm{t}}') \\
\mathcal{T}_{\mathrm{B}} &= |r_1 t_2|^2 + |t_1 r_2|^2 - 2|r_1 t_1 t_2 r_2| \cos(\varphi_\alpha - \varphi_\beta + \theta_{\mathrm{r}} - \theta_{\mathrm{t}}').
\end{aligned}$$

The two transmission probabilities oscillate as a function of the phase difference $\varphi_\alpha - \varphi_\beta$ which can be changed, for example, by changing the optical path length in one of the interferometer arms. The oscillations of the two transmissions will always show a phase difference of π, making sure that $\mathcal{T}_{\mathrm{A}} + \mathcal{T}_{\mathrm{B}} = 1$. This expresses the fact that no photon is lost in the ideal interferometer.

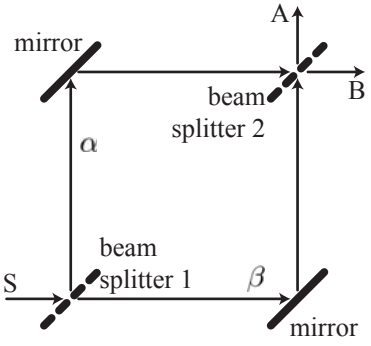


Fig. 16.36 Schematic setup of a Mach–Zehnder interferometer. The photon source is denoted as S; A and B are detectors.

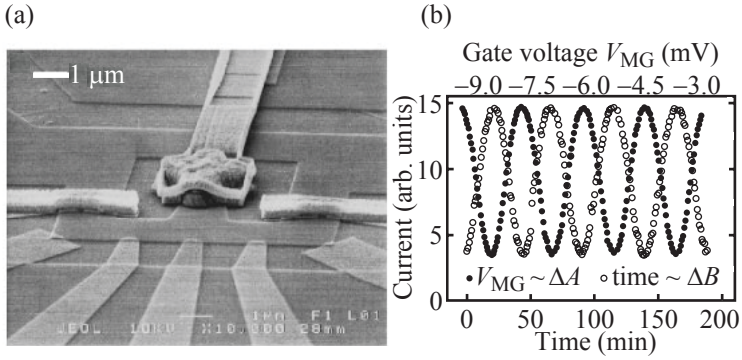


Fig. 16.37 (a) Scanning electron microscope image of the electronic Mach–Zehnder interferometer realized on the basis of a two-dimensional electron gas in Ga[Al]As. (b) Observed interference pattern (see text for details) (Ji *et al.*, 2003. Reprinted by permission from Macmillan Publishers Ltd. Copyright 2003.)

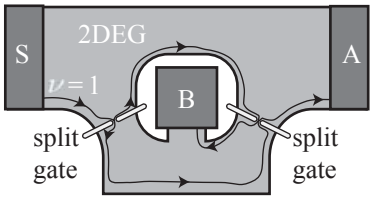


Fig. 16.38 Schematic view of the electronic Mach–Zehnder interferometer. Ohmic contacts are indicated in dark grey.

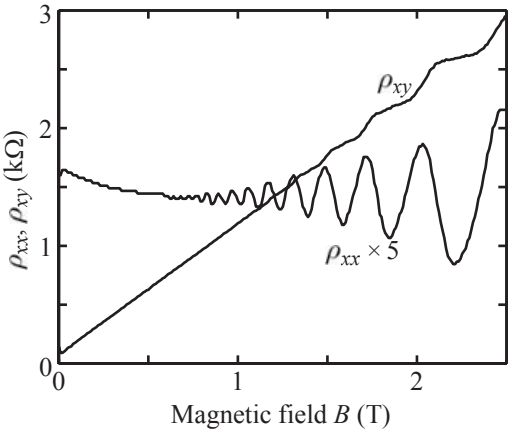
The electronic Mach–Zehnder interferometer has been realized in a sample fabricated from a two-dimensional electron gas in Ga[Al]As (Ji *et al.*, 2003). A scanning electron microscope picture of the corresponding sample is shown in Fig. 16.37(a); a schematic view is presented in Fig. 16.38. Edge states in the quantum Hall regime at filling factor $\nu = 1$ propagating along the sample edges play the role of the light beams in the optical interferometer. Two quantum point contacts are used as beam splitters for quantum Hall edge states by tuning their transmissions $|t_1|^2$ and $|t_2|^2$ to 1/2. The interference patterns were observed at an electronic temperature of 20 mK by applying a voltage between the source S and the drain contacts (A, B) and measuring the transmitted electrons as the currents I_A and I_B. A gate placed near the edge of the sample in one interferometer arm allowed the area enclosed by the two interfering paths to be changed by ΔA, and thereby change the enclosed magnetic flux. This leads to a change of the phase difference $\varphi_\alpha - \varphi_\beta$ by the amount $eB\Delta A/h$. The resulting interference pattern in I_A and I_B is depicted in Fig. 16.37(b). As expected, there is a phase difference of π between the oscillations of the two currents.

Further reading

- Quantum Hall effects: Beenakker and van Houten 1991; Datta 1997; Prange and Girvin 1988; Chakraborty and Pietilainen 1995; Heinonen 1998.
- Papers: von Klitzing *et al.* 1980; Tsui *et al.* 1982; Jain 2000.
- Nobel lecture: Stormer 1999.

Exercises

(16.1) In the figure below you see the magnetic-field-dependent longitudinal resistivity ρ_{xx} and the Hall resistivity ρ_{xy} of a two-dimensional electron gas at the temperature $T = 2.3\,\mathrm{K}$.



(a) Determine the sheet electron density from the Hall effect.

(b) Determine the sheet electron density from the Shubnikov–de Haas oscillations, and compare the resulting value with that determined from the Hall effect.

(c) What is the mobility of the electrons in this sample?

(d) Use the measurement data to estimate the magnetic field at which $\mu B = \omega_c \tau = 1$. What can you directly read from this magnetic field?

(16.2) In the discussion of the Drude model, in eq. (10.50) we introduced the Drude scattering time

$$\frac{\hbar}{\tau_0(E)} = n_i \frac{m^\star}{2\pi\hbar^2} \int_0^{2\pi} d\varphi \left\langle \left| v^{(i)}(\mathbf{q}) \right|^2 \right\rangle_{\mathrm{imp}} (1-\cos\varphi),$$

whereas in the discussion of the Landau level broadening we used the quantum lifetime in eq. (16.6),

$$\frac{\hbar}{\tau_\mathrm{q}(E)} = n_i \frac{m^\star}{2\pi\hbar^2} \int_0^{2\pi} d\varphi \left\langle \left| v^{(i)}(\mathbf{q}) \right|^2 \right\rangle_{\mathrm{imp}},$$

where $|q| = \sqrt{2k_\mathrm{F}^2(1-\cos\varphi)}$.

(a) Discuss the differences between these two scattering times and their physical meaning.

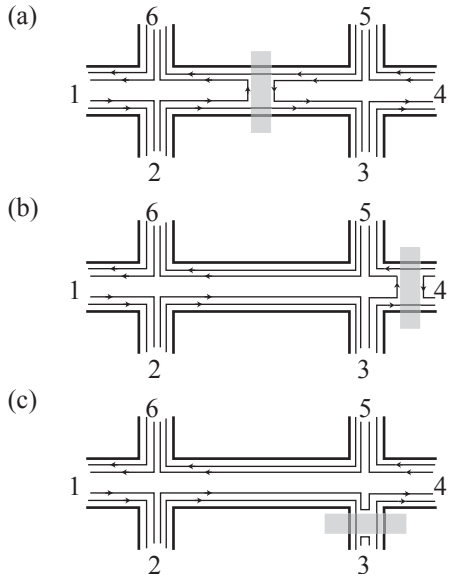
(b) Show that for short-range scattering potentials the ratio $\tau_0(E_\mathrm{F})/\tau_\mathrm{q}(E_\mathrm{F}) \approx 1$. How does this ratio change for long-range scattering potentials?

(16.3) In this problem you consider the degeneracy of Landau levels in a two-dimensional electron gas with a magnetic field applied perpendicular to the plane. Let the sheet electron density of the electron gas be n_s.

(a) Calculate the magnetic field for which there is exactly one magnetic flux quantum h/e per electron. Show that this magnetic field corresponds to filling factor $\nu = 1$.

(b) Applying a magnetic field does not change the sheet density n_s in the electron gas. Calculate the degeneracy of a Landau level by assuming that the two-dimensional density of states at zero magnetic field 'contracts' into Landau levels at finite magnetic field. Hint: neglect the Zeeman splitting of Landau levels, but take spin degeneracy into account.

(16.4) Consider a two-dimensional electron gas in Hall bar geometry in a magnetic field perpendicular to the plane of the electron gas at Landau level filling factor $\nu = 2$. A suitable voltage applied to the metallic top gate shown in gray can be used to reflect individual edge channels as indicated in the figure below. Current is driven from contacts 1 to 4; all other contacts are used for voltage measurements. Compare the two-terminal resistance $R_{14,14}$, the two longitudinal resistances $R_{14,23}$ and $R_{14,65}$, and the two Hall resistances $R_{14,26}$ and $R_{14,35}$ for the cases depicted in Figures (a)–(c) and the generic case, when the gate is not energized.



(16.5) In this problem you reconsider your understanding of the fractional quantum Hall effect.

(a) What are the experimental boundary conditions for the observation of the fractional quantum Hall effect? Discuss measurement temperature, electron mobility, and the magnetic field strength.

(b) Determine the elastic mean free path for electrons in a sample with the mobility $\mu = 30 \times 10^6\,\mathrm{cm}^2/\mathrm{Vs}$. What are the consequences of the result for the experiment?

(c) Why do the filling factor fractions at which the fractional quantum Hall effect occurs have almost exclusively odd denominators?

(d) Discuss in which cases 'composite fermions' or 'composite bosons' are more appropriate descriptions of the physics at particular filling factors.

(e) Do interaction effects have any importance for the integer quantum Hall effect?

(f) Which experiments may be used to prove the existence of a Fermi surface for quasiparticles?

(16.6) Consider a very high mobility two-dimensional electron gas subject to a strong magnetic field normal to the plane driving the system into a region of filling factor $\nu < 1$. The electronic system is assumed to be fully spin-polarized. In this problem we consider the joint motion of two interacting electrons which is governed by the hamiltonian

$$H = \sum_{i=1}^{2} \frac{[\mathbf{p}_i + |e|\mathbf{A}(\mathbf{r}_i)]^2}{2m^\star} + \frac{e^2}{4\pi\epsilon\epsilon_0|\mathbf{r}_1 - \mathbf{r}_2|}.$$

You will now work on a plausibility argument as to why fractional filling factors with odd denominators lead to states with reduced energy.

(a) Show that the hamiltonian can be written as the sum of a part describing the center of mass motion and a part describing the relative motion of the two electrons. The center of mass coordinate is $\boldsymbol{R} = (\mathbf{r}_1 + \mathbf{r}_2)/2$, and the relative coordinate is $\mathbf{r} = \mathbf{r_1} - \mathbf{r}_2$. What is the energy spectrum of the center of mass hamiltonian?

(b) Express the hamiltonian of the relative motion in cylindrical coordinates and introduce a relative angular momentum quantum number m. Discuss the physical meaning of this relative angular momentum quantum number for the motion of the particles.

(c) What implication does the symmetry of the fermionic two-particle wave function upon particle exchange have for the possible relative angular momentum values, given that the two electrons have the same spin?

(d) Although we cannot solve the radial equation analytically, an estimate for the average separation $\bar{\rho}$ of the two electrons can be made by inspection of the hamiltonian. Justify the result

$$\bar{\rho} = 2l_\mathrm{c}\sqrt{|m|},$$

where $l_\mathrm{c} = \sqrt{\hbar/eB}$ is the magnetic length.

(e) A minimum for the energy of an interacting two-dimensional electron gas in a magnetic field is expected, when the average electronic separation

$$\tilde{\rho} = \sqrt{\frac{2}{\pi n_\mathrm{s}}}$$

given by the two-dimensional electron density n_s is commensurate with $\bar{\rho}$. At which fractional filling factors is this commensurability condition fulfilled? What are the three largest possible filling factors?

Interaction effects in diffusive two-dimensional electron transport

17

In previous chapters we have seen that electron–electron interaction effects can have observable consequences. Examples were: screening effects in the two-dimensional electron gas leading to Friedel oscillations of the density (section 9.5), the self-consistent reconstruction of edge channels in the quantum Hall regime (section 16.3.5), and the fractional quantum Hall effect (section 16.4). In this chapter we discuss the effects of electron–electron interaction in low-field magnetotransport properties. This theory is relevant for the diffusive transport of electrons in two-dimensional systems, because the scattering potentials seen by an individual electron moving through the electron gas are screened by the sea of the other electrons in the system. The efficiency of this screening depends on temperature and therefore leads to a temperature dependence of the Drude conductivity. We will see that this temperature dependence is directly related to the occurrence of Friedel oscillations around a scattering center. Multiple scattering at scattering centers and Friedel oscillations will then—similar to the weak localization effect—lead to a logarithmic quantum correction of the Drude conductivity.

17.1 Influence of screening on the Drude conductivity

When we calculated the conductivity from Boltzmann's equation we found the expression (10.36) which we write here for zero magnetic field in the form

$$\sigma_{xx}(T) = \frac{ne^2}{m^\star} \int \frac{dE}{E_F} \frac{E\tau_0(E)}{4k_\mathrm{B}T\cosh^2[(E-\mu)/2k_\mathrm{B}T]}. \tag{17.1}$$

This expression corresponds to the Drude conductivity $\sigma_{xx} = ne^2\langle\tau\rangle/m^\star$, if we use the definition

$$\langle\tau\rangle = \int \frac{dE}{E_F} \frac{E\tau_0(E)}{4k_\mathrm{B}T\cosh^2[(E-\mu)/2k_\mathrm{B}T]} \tag{17.2}$$

for the mean scattering time. For impurity scattering, the energy-dependent scattering time is determined by Fermi's golden rule result

(10.51) according to which

$$\frac{\hbar}{\tau_0(E)} = n_i \frac{m^\star}{2\pi\hbar^2} \int_0^{2\pi} d\varphi \left\langle \frac{\left| v_{\mathrm{ext}}^{(i)}(\mathbf{q}) \right|^2}{\varepsilon^2(q,\mu,T)} \right\rangle_{\mathrm{imp}} (1-\cos\varphi). \tag{17.3}$$

Writing the results of the Drude theory in this way, we can see that the temperature dependence of the conductivity is caused by different sources. It results from the energy averaging in eq. (17.1), and it is also influenced by the explicit temperature dependence of Lindhards dielectric function which has its origin in a similar energy averaging procedure in eq. (9.8). The energy averaging is weighted by the $\cosh^{-2}$-derivative of the Fermi function symmetrically around the chemical potential μ. The latter itself depends on temperature via

$$\mu(T) = k_{\mathrm{B}} T \ln\left(e^{E_F/k_{\mathrm{B}}T} - 1\right),$$

which again has consequences for the energy averaging. This effect is, however, weak for the degenerate case with $k_{\mathrm{B}}T \ll E_{\mathrm{F}}$ considered here, and will therefore be neglected.

The result of the energy averaging depends on the curvature at $E = \mu$ of the remaining integrand because the derivative of the Fermi function is symmetric around μ. The polarization function $\Pi(q,\mu,T)$ entering the dielectric constant is at this energy convex, i.e., $\Pi(q,\mu,T)$ decreases with increasing temperature (see Fig. 9.4). As a result the value of the dielectric function decreases with increasing temperature for the q-values of interest, i.e., the scattering rate increases and the conductivity decreases with increasing temperature. This behavior is called *metallic*, because $d\sigma/dT < 0$ is typical of metals at low temperatures (phonons).

The behavior of the integrand $E\tau_0(E)$ in eq. (17.1) is different: $\tau_0(E)$ typically grows with E (see the discussion about the denisty dependence of the scattering rate on page 169), therefore $E\tau_0(E)$ is concave, and the average scattering time has the tendency to increase with temperature. As a result we find $d\sigma/dT > 0$, i.e., the behavior is called *insulating*. The total temperature dependence of the conductivity at low temperatures is therefore determined by a competition between temperature-dependent screening and the energy averaging of the scattering time. Materials in which large angle scattering dominates, e.g., those with charged impurities close to the electron gas, will exhibit a strong influence of screening on the temperature dependence, because $\varepsilon(q,\mu,T)$ changes with temperature mostly at $q \approx 2k_F$. In contrast, samples with dominant small angle scattering, such as high quality heterostructures with remote doping, will be dominated by the temperature dependence resulting from energy averaging of the scattering time.

We have seen that temperature-dependent screening can lead to a metallic temperature dependence of the conductivity in two-dimensional electron gases. Microscopically this effect can be interpreted as a consequence of scattering at Friedel oscillations arising around a single impurity (Zala *et al.*, 2001). The process is schematically depicted in

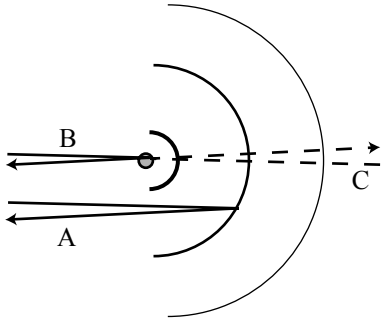


Fig. 17.1 Scattering of electrons at Friedel oscillations forming around a single impurity. Interference between paths A and B is always constructive leading to an enhancement of backscattering. (Reprinted with permission from Zala *et al.*, 2001. Copyright 2001 by the American Physical Society.)

Fig. 17.1. Paths A and B interfere always constructively, thereby increasing backscattering compared to the classical result, because the Friedel-oscillations have a wavelength of exactly $\lambda_F/2$.

Following our discussion of Friedel oscillations in section 9.5 we can express the scattering potential as the sum of a Thomas–Fermi contribution $V_{\mathrm{TF}}(q)$ and the temperature-dependent contribution $V_{\mathrm{friedel}}(q,\mu,T)$ of the Friedel oscillations. The total scattering rate can then be written as

$$\frac{\hbar}{\tau(E)} = n_i \frac{m^\star}{2\pi\hbar^2} \int_0^{2\pi} d\varphi \, |V_{\mathrm{TF}}(q) + V_{\mathrm{friedel}}(q,\mu,T)|^2 \, (1-\cos\varphi).$$

When we square the sum of the two Fourier transformed potentials, we obtain the Thomas–Fermi contribution $|V_{\mathrm{TF}}(q)|^2$, the interference term $2V_{\mathrm{TF}}(q)V_{\mathrm{friedel}}(q)$, which describes the interfering paths A and B introduced in Fig. 17.2, and the contribution $|V_{\mathrm{friedel}}(q,\mu,T)|^2$ describing scattering at Friedel oscillations alone. This last term will be small and therefore negligible, considering the fact that the Friedel oscillations are small compared to the Thomas–Fermi contribution. We are therefore interested in the interference term which is shown in Fig. 17.2 for $T = 0$. For energies $E < E_{\mathrm{F}}$, we have $q < 2k_{\mathrm{F}}$, and the matrix element is zero. Therefore, there is no contribution to the scattering rate for these energies. In contrast, for energies $E \geq E_{\mathrm{F}}$ there is a finite contribution. Figure 17.3 shows the corresponding energy-dependent scattering rate. It can be shown that the total scattering rate at $T = 0$ and close to $E = E_{\mathrm{F}}$ is then given by

$$\frac{\hbar}{\tau} = \frac{\hbar}{\tau_0(E)} + K\frac{E-E_F}{E_F}\Theta(E-E_F),$$

where $\Theta(E)$ is the Heaviside step function and K is a constant. This additional energy dependence of the scattering time results after energy averaging with the derivative of the Fermi function in eq. (17.2) in a linear temperature dependence of the conductivity. In this line of reasoning, we have, however, neglected the fact that the amplitude of the Friedel oscillations also depends on temperature. Taking this effect into account as well, the temperature dependence remains linear and acquires the form (Gold and Dolgopolov, 1986)

$$\sigma_{xx}(T) = \frac{ne^2\langle\tau(T=0)\rangle}{m^\star}\left(1 - C\frac{k_{\mathrm{B}}T}{E_F} + \cdots\right), \qquad (17.4)$$

where C is a constant that depends on the specific scattering mechanism. In this sense, the temperature dependence of the conductivity at low temperatures is an interplay of interactions and interference, if temperature-dependent screening dominates.

Of course, this metallic temperature dependence competes with the effect of weak localization. At sufficiently low temperatures the localizing effects take over. This is shown in Fig. 17.4 for a p-SiGe two-dimensional hole gas. At the highest magnetic fields where weak localization effects

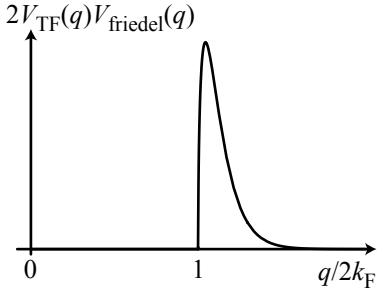


Fig. 17.2 Matrix element of the interference term of scattering at the Thomas–Fermi screened potential and the Friedel oscillations.

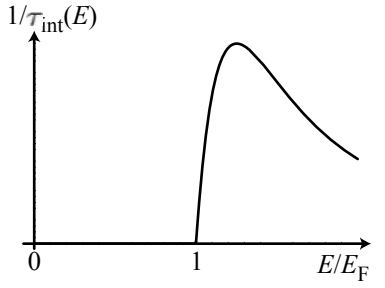


Fig. 17.3 Energy dependence of the scattering rate for scattering at Friedel oscillations in lowest (first) order.

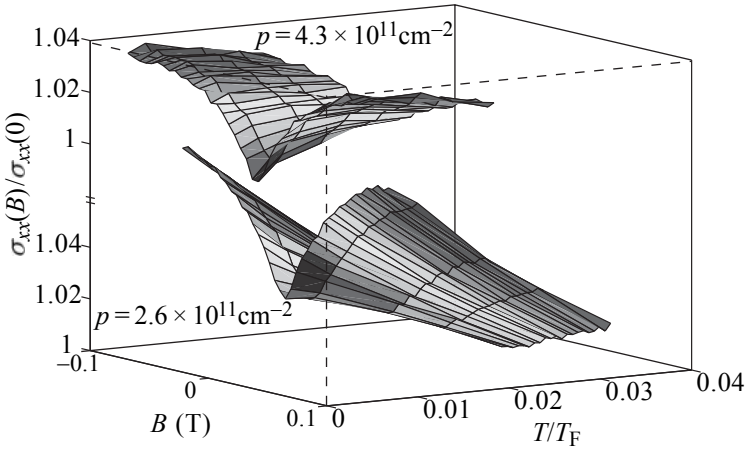


Fig. 17.4 Temperature and magnetic field dependence of a two-dimensional hole gas in a p-SiGe quantum well for two different charge carrier concentrations. The data show the interplay between temperature-dependent screening and weak localization. (Reprinted with permission from Senz *et al.*, 2000*b*. Copyright 2000 by the American Physical Society.)

are suppressed, the conductance increases for both charge carrier concentrations. At the magnetic field $B = 0$, however, the weak localization maximum acts more and more strongly against the metallic behavior as the temperature is lowered.

17.2 Quantum corrections of the Drude conductivity

With the microscopic picture of temperature-dependent screening as scattering at Friedel oscillations of *single* scattering sites, it is obvious to ask, whether *multiple scattering* at impurity sites and Friedel oscillations can also have an influence on the conductance at low temperatures. The theoretical answer to this question was given in the 1980s within a diagrammatic theory for low temperatures $k_\text{B}T \ll \hbar/\tau$ (Altshuler and Aronov, 1985). The intuitive picture shown in Fig. 17.5 was developed later (Zala *et al.*, 2001). As in the case of the weak localization, a logarithmic correction for the conductivity results:

$$\delta\sigma_\text{I} = -\frac{e^2}{\pi h}\left(1 + \frac{3}{4}\tilde{F}_\sigma\right)\ln\left(\frac{\hbar/\tau}{k_\text{B}T}\right) \quad \text{for } k_\text{B}T \ll \hbar/\tau \tag{17.5}$$

Here, $\tilde{F}_\sigma$ is an interaction parameter. With decreasing temperature the conductance decreases, i.e., this interaction correction to the conductivity helps to localize the electronic system for $T \to 0$. All the different corrections to the Drude conductivity—weak localization, temperature-dependent screening, and interaction correction—are at low temperatures additive in first order.

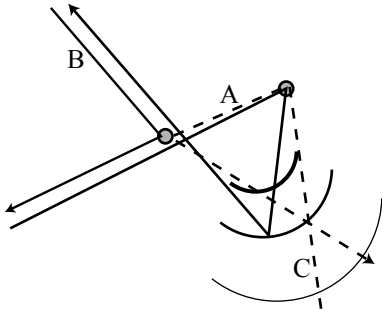


Fig. 17.5 Multiple scattering of electrons at impurity sites and at Friedel oscillations. (Reprinted with permission from Zala *et al.*, 2001. Copyright 2001 by the American Physical Society.)

Experimentally one can distinguish the logarithmic interaction correction from the weak localization correction via their magnetic field dependencies. The interaction corrections do not depend on magnetic fields if these are sufficiently small, whereas the weak localization effects are suppressed by small fields. Another possibility for distinguishing the

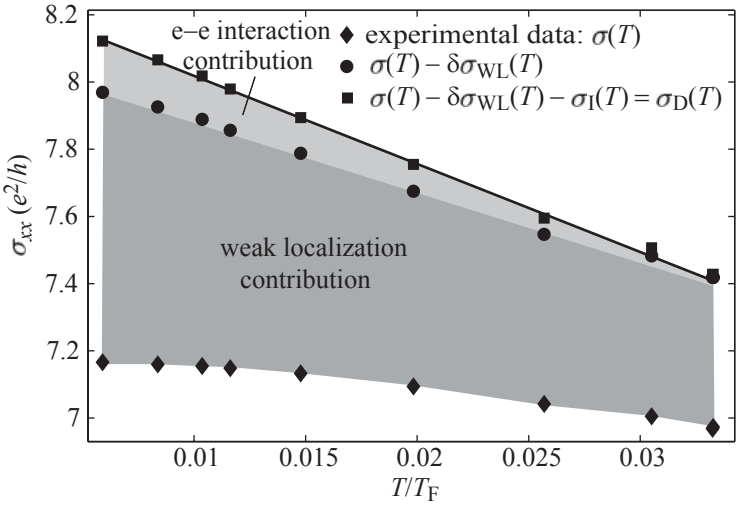


Fig. 17.6 The different contributions to the conductivity of a *p*-SiGe sample at low temperatures. The Drude conductivity including temperature-dependent screening is indicated with black squares. Adding the logarithmic interaction correction one obtains the filled circles. Further addition of the weak localization correction leads to the black diamonds, i.e., the measured conductivity. (Reprinted with permission from Senz *et al.*, 2000*b*. Copyright 2000 by the American Physical Society.)

two is the measurement of the Hall resistivity. The weak localization effect does not alter the Drude result for the Hall effect. In contrast, the interaction correction leads to a temperature dependence of the Hall slope (for experimental details, see, e.g., Glew *et al.*, 1981, or Senz *et al.*, 2000*b*). Figure 17.6 shows the different contributions to the conductance of a *p*-SiGe sample.

Further reading

- Papers: Glew *et al.* 1981; Zala *et al.* 2001; Senz *et al.* 2000*b*.

Exercises

(17.1) In this exercise you reconsider your understanding of interaction effects in two-dimensional electron gases.

(a) Name at least three effects of electron–electron interactions in two-dimensional electron transport that you have learned about so far in this book.

(b) Discuss qualitatively what role interaction effects could play in transitions from the weakly localized to the strongly localized regimes.

(c) What property do electrons in a heterostructure need to have for multiple scattering from impurities and Friedel oscillations to be particularly important? What other quantum corrections to the conductivity will arise under these circumstances?

This page intentionally left blank

Quantum dots

18

18.1 Coulomb-blockade effect in quantum dots

In the nanostructures that we have been discussing so far in this book, the electron–electron interaction has played a minor role and could be taken into account perturbatively, if necessary. In this chapter we are going to discuss the physics of quantum dots. In these systems, Coulomb interactions can play a dominant role compared to other energy scales. In particular, we will discuss the Coulomb blockade effect, which is one of the fundamental transport phenomena in semiconductor nanostructures. Quantum dots also differ from other nanostructures discussed earlier, because they are very weakly coupled to their environment, i.e., to source and drain leads, or to lattice vibrations and thermal electromagnetic radiation at low temperatures. These structures confine electrons or holes in regions of space small enough to make their quantum mechanical energy levels observable.

18.1.1 Phenomenology

The top left inset of Fig. 18.1 shows an image taken with a scanning electron microscope of a quantum dot structure defined with the split-gate technique. The two outer pairs of gates are used to deplete the electron gas under the surface and to form quantum point contacts. These two quantum point contacts are connected in series. Between them there are two further gate electrodes, the so-called plunger gates. These allow us to tune the electron density in the region between the quantum point contacts. In the experiment a small voltage of the order of $10\,\mu\mathrm{V}$ is applied for linear conductance measurements at a temperature of about $100\,\mathrm{mK}$ between the source and the drain contact which are formed by the two-dimensional electron gas outside the dot structure. The voltage results in a current through the quantum dot which can be controlled with the plunger gates. At such small source–drain voltages V_{SD} the current I is linearly related to this voltage and therefore the linear conductance is given by $G = I/V_{\mathrm{SD}}$. If larger source–drain voltages are applied the $I(V_{\mathrm{SD}})$ curves become nonlinear and one often measures not only $I(V_{\mathrm{SD}})$, but also the differential conductance dI/dV_{SD} as a function of a constant V_{SD}. This is achieved by superimposing a small (e.g., $10\,\mu\mathrm{V}$) low-frequency (e.g., between 10 and 30 Hz) alternating δV_{SD} and monitoring the resulting alternating current component δI. The differ-

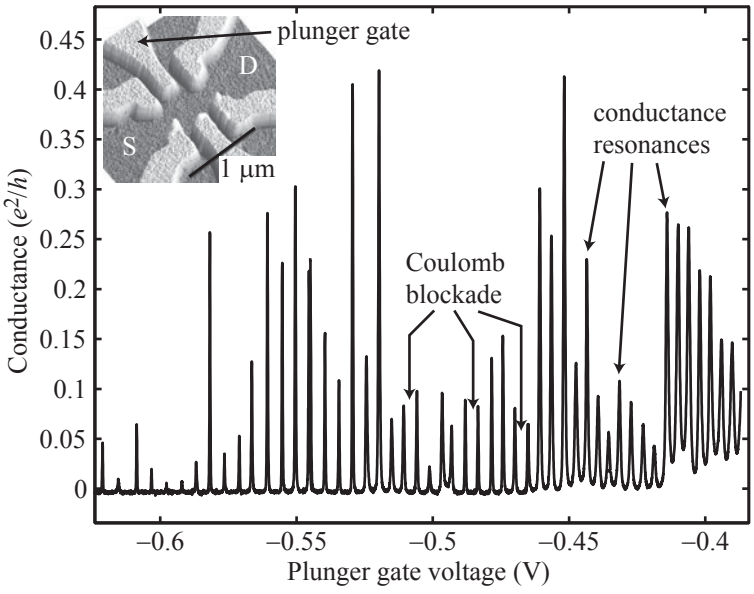


Fig. 18.1 Quantum dot linear conductance as a function of the plunger gate voltage. Inset: split gate structure used to define the quantum dot. The conductance is measured by applying a small voltage between source (S) and drain (D), and recording the source drain current. (Reprinted with permission from Lindemann *et al.*, 2002. Copyright 2002 by the American Physical Society.)

ential conductance is then given by $dI/dV_{\rm SD} = \delta I/\delta V_{\rm SD}$.

Conductance resonances. The measured curve in Fig. 18.1 shows the linear conductance (the measured current divided by the applied voltage) of this quantum dot as a function of the plunger gate voltage. The quantum dot conductance shows sharp resonances. Between them the current is zero within measurement accuracy. We will see below that the Coulomb interaction between electrons in the quantum dot is crucial for the understanding of this measurement. Therefore it is called the *Coulomb blockade effect*.

Nonlinear current–voltage characteristics. The Coulomb blockade effect also manifests itself in measurements of nonlinear $I(V_{\rm SD})$ characteristics of quantum dots. Figure 18.2 shows two $I(V_{\rm SD})$ curves, one of which ('on resonance') was measured at the plunger gate voltage of a conductance peak in Fig. 18.1, the other one in a valley between two conductance resonances ('off resonance'). The trace taken on a conductance resonance increases linearly with the bias voltage in the region $|V_{\rm SD}| < 4k_{\rm B}T$. In contrast, the current measured between resonances is suppressed for source–drain voltages that are significantly larger than $4k_{\rm B}T$, rising only at very large applied bias voltages. In regions of suppressed current in Figs. 18.1 and 18.2 the quantum dot is said to be in the Coulomb blockade.

Coulomb blockade diamonds. The measurements of the differential conductance $dI/dV_{\rm SD}$ as a function of the plunger gate $V_{\rm pg}$ and as a function of the source–drain voltage $V_{\rm SD}$ can be combined resulting in the

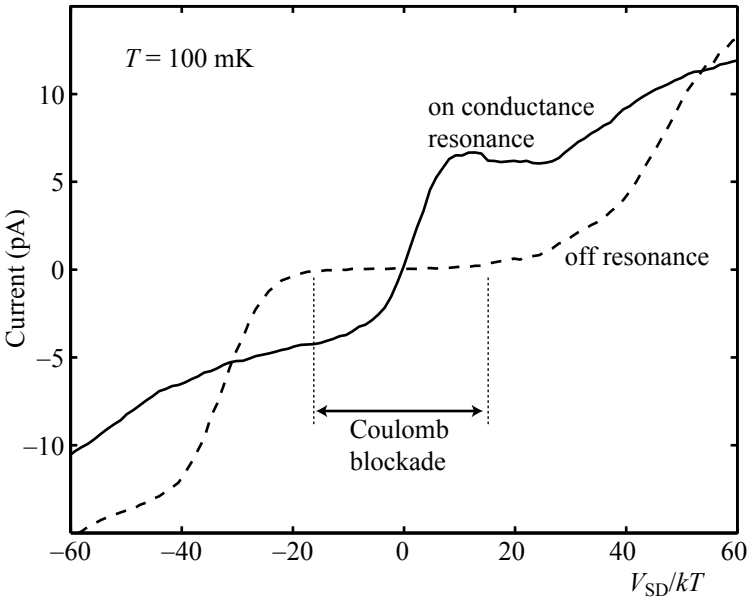


Fig. 18.2 Measured $I(V_{SD})$ trace of a quantum dot on a conductance resonance (solid line) and off resonance (dashed line). In the latter case the current is suppressed at low V_{SD} (Coulomb blockade).

measurement of the so-called Coulomb blockade diamonds. Figure 18.3 shows the result of such a measurement in the $V_{pg} - V_{SD}$ plane. The differential conductance is represented by the gray scale, where black represents zero and white finite positive values.

The bright spots along the line $V_{SD} = 0$ are the conductance peaks of linear transport, as depicted in Fig. 18.1. Between these conductance peaks there are diamond shaped black regions in which electron transport is completely suppressed as a result of the Coulomb-blockade effect.

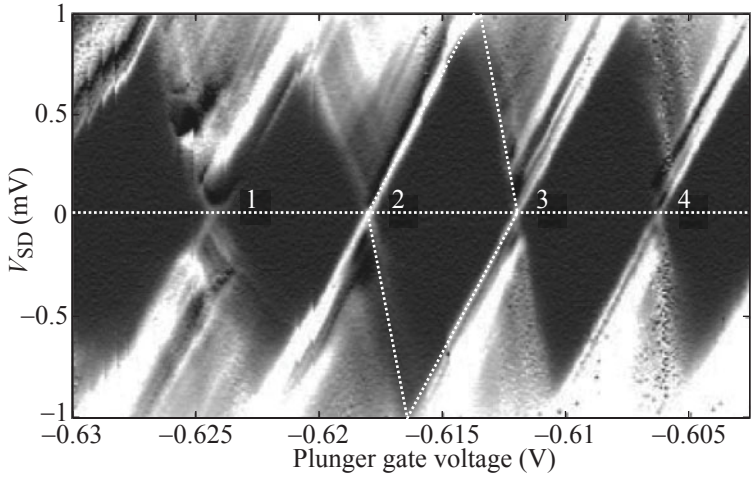


Fig. 18.3 Coulomb blockade diamonds measured on a quantum dot realized on a parabolic quantum well. The grayscale represents the differential conductance dI/dV_{SD} with zero encoded in black. Numbers 1–4 label zero source–drain voltage conductance peaks. One particular diamond is highlighted with a dotted boundary, the horizontal dotted line indicates $V_{SD} = 0$. (Reprinted with permission from Lindemann *et al.*, 2002. Copyright 2002 by the American Physical Society.)

18.1.2 Experiments demonstrating the quantization of charge on the quantum dot

The importance of the Coulomb interaction and the quantization of charge in units of $-|e|$ in quantum dots can be demonstrated in experiments in which an additional quantum point contact is fabricated close to the dot without direct tunneling coupling between the dot and the additional point contact, but only capacitive coupling. As an alternative to the additional quantum point contact, a second quantum dot can be used. These additional devices serve as sensors of the charge residing on the quantum dot. The basic idea is that additional charge sucked into the quantum dot using the plunger gate will change the potential landscape in the environment around the dot. This change can be detected with the additional device. It turns out that the change is not continuous like the change of charge on a macroscopic capacitor, but occurs in steps of one elementary charge $-|e|$. Therefore these experiments establish the quantization of charge on the dot and demonstrate the possibility of controlling individual electrons in quantum dot devices.

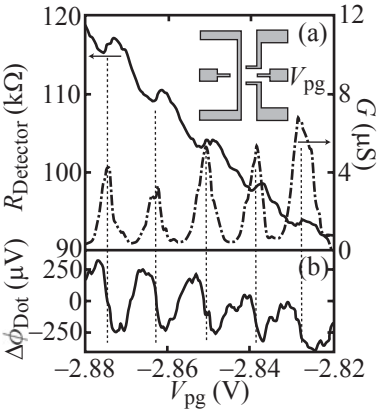


Fig. 18.4 (a) Right axis: Quantum dot conductance (dash-dotted) as a function of the plunger gate voltage V_{pg} showing conductance resonances. Left axis: Resistance of the quantum point contact detector as a function of the same plunger gate voltage. Inset: split gate structure used for defining the devices. (b) Detector signal where the dependence of the detector sensitivity on the plunger gate voltage has been calibrated out. (Reprinted with permission from Field *et al.*, 1993. Copyright 1993 by the American Physical Society.)

Measurement of charge quantization in a quantum dot using a quantum point contact. At low temperatures, a split-gate defined quantum point contact shows quantized conductance as a function of the voltage applied to the split-gate. At the transition between quantized conductance plateaus (e.g., at a conductance value of about e^2/h), the conductance is very sensitive to small changes of the gate voltage, or other local electrostatic potentials. The arrangement depicted schematically in the inset of Fig. 18.4 allows us to observe the stepwise charging of the quantum dot with single electrons using the quantum point contact as a charge detector. The schematic picture shows a top view onto gates (gray) which were patterned on top of a Ga[Al]As heterostructure. The quantum dot is defined by depleting the electron gas under the gates. The detector quantum point contact forms at a distance of about 300 nm from the quantum dot center. There is no tunneling coupling between this quantum point contact and the quantum dot. The coupling between the two devices is purely capacitive. The measured curve in Fig. 18.4(a) shows the conductance of the quantum dot with a series of conductance resonances (right axis). The other curve belonging to the left axis shows the resistance of the quantum point contact. This resistance shows rapid increases whenever the quantum dot conductance goes through a resonance (see dashed lines). This demonstrates that the electrostatic potential in the quantum dot changes stepwise, whenever a quantum $-|e|$ of charge is added to the quantum dot, i.e., whenever a single electron is added. Adding charge to the dot is related to current flow into the dot. The curve in Fig. 18.4(b) is the signal of the detector that corresponds to the curve in (a), but the vertical axis has been calibrated to represent the electrostatic potential in the quantum dot. The calibration emphasizes the oscillatory change of the potential. The potential rises with increasing plunger gate voltage. However, each

added electron lowers the potential again as a result of its own charge.

Current through the quantum dot is not really necessary for the observation of the quantized charge. It is sufficient if the quantum dot is connected via a tunneling contact to a single electron reservoir. Increasing the plunger gate voltage will then allow one electron after the other to be added to the quantum dot. If the bandwidth of the quantum point contact detector circuit is made sufficiently fast, i.e., larger than the tunneling rate, single electron hopping between the quantum dot and the lead can be observed in real time (Schleser *et al.*, 2004). A typical time trace of such a measurement is shown in Fig. 18.5. Whenever an electron hops into the dot the quantum point contact current jumps down. When the electron hops out again, the original current value is restored. Since electron hopping is a quantum mechanical tunneling process, it occurs randomly in time. The time that the electron spends inside the quantum dot has an exponentially decaying distribution function.

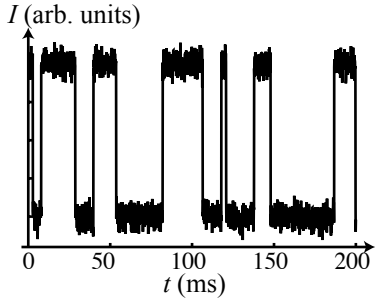


Fig. 18.5 Current through a detector quantum point contact switching in time as a result of charging and discharging of a nearby quantum dot with a single electron.

Dependence of the charge quantization on the tunneling coupling. Figure 18.1 shows the reaction of the quantum dot conductance when the plunger gate voltage is changed. At the steep slopes of a conductance resonance, the conductance is extremely sensitive to very small changes of the applied voltage. This property can be exploited for using a quantum dot as a charge detector which measures small changes in the charge state of its neighborhood. Figure 18.6 shows an experimental arrangement in which a quantum dot (Box) is placed next to a quantum dot charge detector (Elect). The detector is tuned to the steep slope of a conductance resonance using the plunger gate voltage V_{e}. The number of electrons in the dot to be measured is varied with the plunger gate V_{pg}. The result of this measurement is shown in Fig. 18.7 for varying coupling strengths between the 'Box' and its leads. For the weakest coupling (a) charge quantization is very well expressed which appears as the series of very sharp steps in the detector current. With increasing coupling the steps smear out more and more until they have disappeared completely once the coupling conductance has reached a value of about twice the conductance quantum $2e^2/h$. This result shows that the charge on the quantum dot 'box' is only quantized if the tunneling resistance to source and drain far exceeds half the resistance quantum $h/2e^2$.

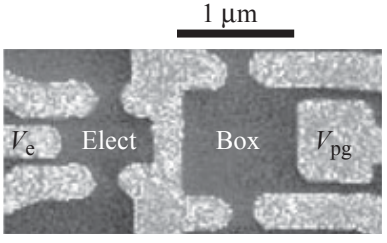


Fig. 18.6 Scanning electron micrograph of a circuit in which one quantum dot (Elect) is employed as a charge detector for the charge state of the neighboring quantum dot (Box). The structure is realized on a two-dimensional electron gas, bright regions represent metallic electrodes on the sample surface. (Reprinted with permission from Duncan *et al.*, 1999. Copyright 1999, American Institute of Physics.)

18.1.3 Energy scales

The Coulomb blockade effect in quantum dots is characterized by an interplay of several energy scales which we will estimate in the following. Coulomb blockade in the strict sense arises only if the Coulomb energy dominates over all other energy scales.

Coulomb energy. The size of the Coulomb interaction energy between electrons on a quantum dot island can be estimated by determining its capacitive charging energy. For the sake of the argument we choose a quantum dot made from a two-dimensional electron gas. We regard it as

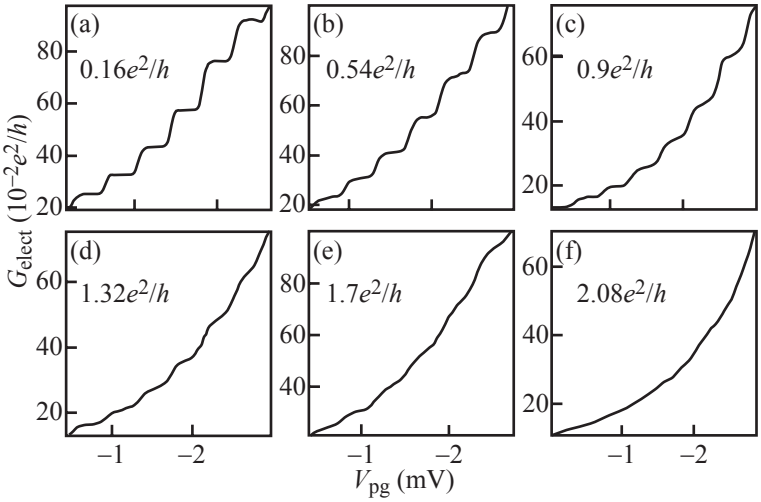


Fig. 18.7 Charge quantization in a quantum dot as a function of the tunneling coupling to source and drain. The strength of the coupling is increased from (a) to (f). (Reprinted with permission from Duncan *et al.*, 1999. Copyright 1999, American Institute of Physics.)

a metallic disc of radius r carrying the charge $-|e|N$ which is embedded in a homogeneous dielectric material with relative dielectric constant ε. The self-capacitance C of such a disc is $C = 8\varepsilon\varepsilon_0 r$. Inserting $\varepsilon = 13$ and $r = 100\,\mathrm{nm}$ gives $C = 92\,\mathrm{aF}$. The electrostatic energy of the island with N electrons is then

$$E_{\rm elstat}(N) = \frac{e^2N^2}{2C} = \frac{e^2N^2}{16\varepsilon\varepsilon_0 r} = \frac{\pi}{2}E^\star_{\rm Ry}\frac{a^\star_{\rm B}}{r}N^2.$$

We see that the electrostatic energy of the island is proportional to the square of the number of electrons on the island and inversely proportional to the size of the island.

Assume that there are already N electrons on the island and we would like to add another electron. The charging energy required is

$$E_{\rm c}(N{+}1) = E_{\rm elstat}(N{+}1) - E_{\rm elstat}(N) = \frac{e^2}{C}(N{+}1/2) \approx \frac{e^2}{C}N = \frac{e^2}{8\varepsilon\varepsilon_0 r}N. \tag{18.1}$$

Here we have assumed $N \gg 1$. This energy scale is proportional to the electron number and inversely proportional to the island size.

Traditionally, however, in quantum dot physics the term charging energy is used for the difference

$$\Delta E_{\rm c} = E_{\rm c}(N+1) - E_{\rm c}(N) = \frac{e^2}{C} = \frac{e^2}{8\varepsilon\varepsilon_0 r}. \tag{18.2}$$

Taking the above numerical values the characteristic energy scale for charging the island is therefore given by $\Delta E_{\rm c} = e^2/C \approx 1.7\,\mathrm{meV}$, corresponding to a temperature of about 19 K.

Confinement energy. We compare the charging energy scale with the confinement energy of quantum states on the island. Assuming again

a quantum dot realized in a two-dimensional electron gas with Fermi energy E_F we can estimate the number of electrons on the island to be

$$N = \pi r^2 \frac{m^\star}{\pi\hbar^2} E_\mathrm{F}.$$

This is the dot area πr^2 multiplied with the two-dimensional density of states $m^\star/\pi\hbar^2$ times the Fermi energy E_F. The total energy of this system can be estimated to be

$$E_\mathrm{conf}(N) = \int_0^{E_\mathrm{F}} E\pi r^2 \frac{m^\star}{\pi\hbar^2} dE = \pi r^2 \frac{m^\star}{2\pi\hbar^2} E_\mathrm{F}^2 = \frac{\hbar^2}{2m^\star r^2} N^2 = E_\mathrm{Ry}^\star \left(\frac{a_\mathrm{B}^\star}{r}\right)^2 N^2,$$

where we have expressed the Fermi energy with the help of the above equation as the electron number. We can see that the confinement energy is also proportional to the square of the electron number, but it scales inversely proportional to the square of the quantum dot size. We mention here that this result is only valid for systems with a parabolic dispersion relation, because the latter determines the expression for the density of states. For example, the same type of argument leads to $1/r$ scaling of the confinement energy in graphene quantum dots because the dispersion relation $E(\mathbf{k})$ is linear in $|\mathbf{k}|$ near the relevant K- and K'-points forming the corners of the first Brillouin zone in graphene.

Continuing the discussion of dots in materials with parabolic dispersion relation we find the energy

$$\epsilon(N+1) = E_\mathrm{conf}(N+1) - E_\mathrm{conf}(N) = 2E_\mathrm{Ry} \left(\frac{a_\mathrm{B}^\star}{r}\right)^2 N \qquad (18.3)$$

for adding an additional electron onto a dot already containing N electrons.

The mean spacing between successive energy levels is then given by

$$\Delta = \epsilon(N+1) - \epsilon(N) = 2E_\mathrm{Ry} \left(\frac{a_\mathrm{B}^\star}{r}\right)^2. \qquad (18.4)$$

This energy scale is often called the single-particle level spacing. Assuming a dot radius of 100 nm in a two-dimensional electron gas in GaAs, we obtain an energy of about 110 µeV which is about an order of magnitude smaller than the corresponding charging energy scale in the same system.

Alternatively the single-particle level spacing can be estimated using the model of a quantum mechanical harmonic oscillator for the quantized states in the quantum dot. The ground state of such an oscillator has the spatial extent $2r = \sqrt{\hbar/(m^\star\omega_0)}$, where ω_0 is the frequency of the oscillator. As a result we find the characteristic energy scale $\Delta = \hbar\omega_0 = \hbar^2/(4m^\star r^2)$. For a quantum dot in GaAs with $m^\star = 0.067m_0$ and $r = 100$ nm we obtain about 30 µeV which is of the same order of magnitude as the above result.

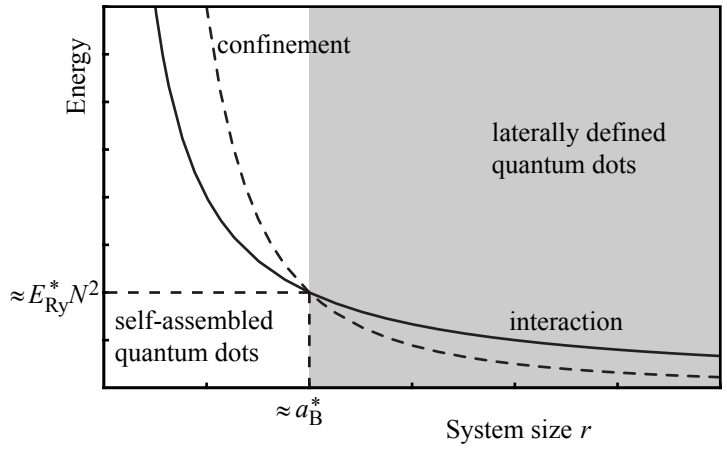


Fig. 18.8 Comparison of Coulomb- and quantization energy in quantum dots. Small dots are dominated by the spatial quantization of states; larger dots are dominated by the Coulomb interaction.

Comparison of Coulomb energy and quantization energy. A comparison of the Coulomb energy and the quantization energy shows that the ratio between the two depends only on the island size (see, e.g., Bryant, 1987, or Brandes *et al.*, 1993). The situation is graphically illustrated in Fig. 18.8. If $r \gg a^\star_{\mathrm{B}}$, the total energy of the quantum dot is dominated by the interaction energy. In contrast, if $r \ll a^\star_{\mathrm{B}}$, the quantization energy is dominant. The crossover between these two regimes is close to the Bohr radius $a^\star_{\mathrm{B}}$. In GaAs it is at about $a^\star_{\mathrm{B}} = 10\,\mathrm{nm}$. For larger islands, like those fabricated by lateral confinement from two-dimensional electron gases in n-GaAs we see that the Coulomb energy is dominant over the quantization energy. This means that we can expect that even in a fully quantum mechanical treatment of such dots, interaction effects will be dominant.

The situation is different, for example, in self-assembled InAs quantum dots. In InAs we have $a^\star_{\mathrm{B}} = 30\,\mathrm{nm}$. Typical dots have a size of 20 nm, i.e., the quantization energy will dominate over the interaction energy.

Source-drain coupling. The quantization of the particle number N on the island is crucial for the observation of the Coulomb blockade effect. The following consideration supports the observation that the coupling strength of the island to source and drain contacts is the important parameter. Charging the island with an additional charge takes the time $\Delta t = R_{\mathrm{t}} C$ which is the RC-time constant of the quantum dot. If we wish to resolve the charging energy $\Delta E_{\mathrm{c}} = e^2/C$ the system will respect Heisenberg's uncertainty relation $\Delta E_{\mathrm{c}} \Delta t > h$ which leads to the condition

$$R_{\mathrm{t}} > \frac{h}{e^2},$$

in agreement with the experiment discussed above. This result means that the tunneling resistance R_{t} of the quantum dot has to be significantly larger than the resistance quantum h/e^2 implying that the quantum point contacts coupling the system to source and drain have to be

deep in the tunneling regime. The result also implies that the uncertainty relation allows the measurement of the charging energy whenever the tunneling coupling is weak enough to quantize the electron number on the island.

Temperature. Another condition for the observation of the Coulomb blockade effect is that the temperature is small compared to the charging energy, i.e.,

$$k_{\mathrm{B}}T \ll \frac{e^2}{C}.$$

In order to fulfill this condition in an experiment, the island must be sufficiently small such that the island capacitance is small. In contrast, macroscopic islands will have $k_{\mathrm{B}}T \gg e^2/C$ for all experimentally accessible temperatures, because the capacitance C is very large.

If the temperature is also small compared to the single-particle level spacing, i.e., if

$$k_{\mathrm{B}}T < \Delta,$$

typically only one quantized energy level contributes to electron transport on a conductance resonance. This is called the single-level transport regime.

If we compare the tunneling coupling Γ of dot states with the temperature scale, we can distinguish the following two scenarios. In the case $k_{\mathrm{B}}T \gg \Gamma$ the conductance resonances will be thermally broadened. In the opposite case of $k_{\mathrm{B}}T \ll \Gamma$, the resonances will be broadened by the tunneling coupling.

18.1.4 Qualitative description

Conductance resonances. On a qualitative level the Coulomb blockade effect can be described in a very intuitive and yet very general way. In a first step we try to understand the occurrence of conductance resonances as a function of plunger gate voltage, as shown in Fig. 18.1. For this purpose we consider the schematic drawing in Fig. 18.9(a). The quantum dot is coupled to a source and drain contact via tunneling barriers. The plunger gate allows us to tune the quantum mechanical energy states in the quantum dot via capacitive coupling. Talking about the quantum mechanical energy states in the quantum dot (which are in most cases hard to work out), we introduce the following notation. As the quantum dot is an almost isolated system, its number of electrons is an integer N. Each N-electron quantum dot will have a ground state (the state with lowest energy), and a sequence of excited states that we label by integer numbers n for convenience. A particular quantum state (N, n) is therefore characterized by the energy $E_N^{(n)}$. We choose $n = 0$ for the ground state and label the states such that $E_N^{n+1} \geq E_N^{(n)}$ for each n. Energy spectra for three different electron numbers on the quantum dot are schematically shown in Fig. 18.10.

It turns out that each energy $E_N^{(n)}$ also depends on the voltage V_{pg} applied between the plunger gate and the quantum dot. In sufficiently

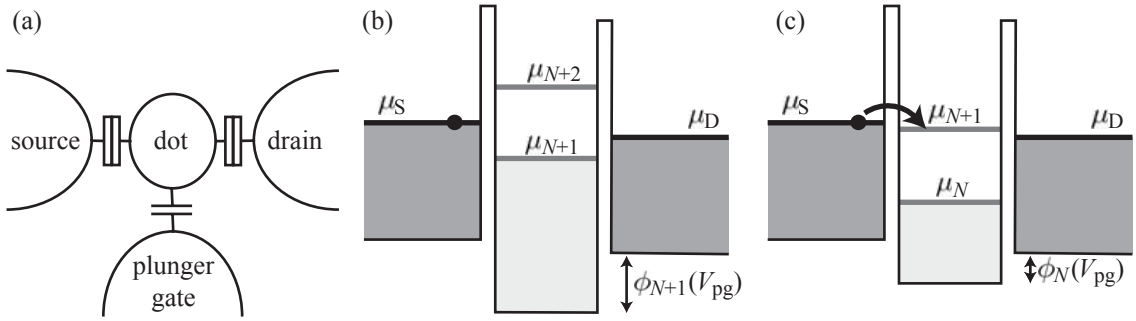


Fig. 18.9 (a) Schematic representation of a quantum dot system with source and drain contacts and a plunger gate. (b) Energy level structure of the system in the Coulomb blockade. (c) Position of the energy levels that allows a current to flow between source and drain if a very small bias voltage is applied.

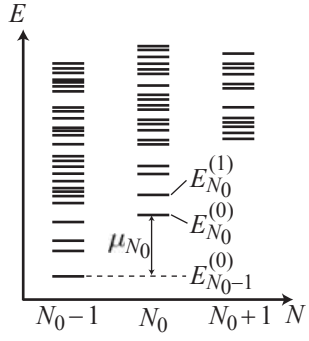


Fig. 18.10 Typical energy spectra of a quantum dot with $N_0 - 1$, N_0, and $N_0 + 1$ electrons.

small ranges of plunger gate voltages around a specific value $V_{\rm pg}^{(0)}$, we can expand

$$E_N^{(n)}(V_{\rm pg}) = E_N^{(n)}(V_{\rm pg}^{(0)}) - |e|N\alpha_{\rm pg}\Delta V_{\rm pg}, \tag{18.5}$$

where the constant $\alpha_{\rm pg}$ is called the *lever arm* of the plunger gate, and $\Delta V_{\rm pg} = V_{\rm pg} - V_{\rm pg}^{(0)}$. In experiments it is found that $\alpha_{\rm pg}$ is in most cases only a very weak function of both N and n such that this dependence can usually be neglected.

Figure 18.9(b) shows the energetic situation in the three subsystems source, drain, and dot. At low temperature the electronic levels in the source (drain) contact are filled from the bottom of the conduction band up to the electrochemical potential $\mu_{\rm S}$ ($\mu_{\rm D}$). In the quantum dot we can also define an electrochemical potential. It describes the energy necessary to add an electron to the dot, given that it is both initially and after the addition in its ground state. For example, if we consider a quantum dot with $N-1$ electrons initially, we define the electrochemical potential for adding the Nth electron as (see also Fig. 18.10)

$$\mu_N(V_{\rm pg}) = E_N^{(0)}(V_{\rm pg}) - E_{N-1}^{(0)}(V_{\rm pg}). \tag{18.6}$$

This energy difference will contain contributions from the electron–electron interaction in the quantum dot, but it may also contain contributions from the confinement energy. Following our simple estimate of the quantum dot energy scales in eqs (18.1) and (18.3) we may identify $\mu_N \approx E_{\rm c} + \epsilon(N)$, where we have neglected any plunger gate voltage dependence. However, this is merely a crude estimate, whereas our model relying on the electrochemical potential definition in eq. (18.6) is much more general. The important message is that—similar to an atom—there is a finite amount of energy needed to add an electron to the quantum dot.

The plunger gate voltage allows us to shift the levels $\mu_N(V_{\rm pg})$ in energy. Combining eqs (18.5) and (18.6) we find for the voltage dependence

of the electrochemical potential in the quantum dot the linear relation

$$\mu_N(V_{\rm pg}) = \mu_N(V_{\rm pg}^{(0)}) - |e|\alpha_{\rm pg}\Delta V_{\rm pg} \tag{18.7}$$

which is independent of the electron number N.

Using the plunger gate voltage we can tune the quantum dot electrochemical potentials into the position shown in Fig. 18.9(c), where we have

$$\mu_{\rm S} \approx \mu_{N+1}(V_{\rm pg}) \approx \mu_{\rm D}.$$

In this case, the energy gain $\mu_{\rm S}$ from removing an electron from the source contact is exactly equal to $\mu_{N+1}(V_{\rm pg})$, the energy required to add an electron to the dot. Once the electron is in the dot, the energy gain $\mu_{N+1}(V_{\rm pg})$ for removing it again is exactly equal to the energy $\mu_{\rm D}$ required to add it to the drain contact. Therefore, elastic electron transport through the quantum dot is possible and the conductance measurement shows a large current (conductance peak, cf., Fig. 18.1) at the respective plunger gate voltages. However, electrons can only tunnel one after another through the dot, because the energy difference $E_{N+2}^{(0)}(V_{\rm pg}) - E_N^{(0)}(V_{\rm pg})$ to add two electrons to the dot at the same time is significantly higher than the energy for a single electron. We therefore talk about *sequential single-electron tunneling*.

The situation of the current blockade is shown in Fig. 18.9(b). At this plunger gate voltage the dot is filled with $N+1$ electrons. In order to fill the $(N+2)$th electron more energy is required than the energy gain from removing an electron from the source contact, i.e.,

$$\mu_{N+2}(V_{\rm pg}) > \mu_{\rm S}, \mu_{\rm D}.$$

The current flow is therefore blocked and we talk about *Coulomb blockade*. This situation corresponds to the $V_{\rm pg}$ regions of suppressed current in Fig. 18.1.

We now want to work out the separation of conductance peaks in plunger gate voltage. To this end we assume that at the Nth conductance resonance $\mu_{\rm S} = \mu_{\rm D} = \mu_N(V_{\rm pg}^{(0)})$ and at the next resonance $\mu_{\rm S} = \mu_{\rm D} = \mu_{N+1}(V_{\rm pg}^{(0)} + \Delta V_{\rm pg}) = \mu_{N+1}(V_{\rm pg}^{(0)}) - |e|\alpha_{\rm pg}\Delta V_{\rm pg}$. Taking the difference between the two equations we find for the separation $\Delta V_{\rm pg}$ of neighboring conductance resonances

$$\Delta V_{\rm pg} = \frac{\mu_{N+1}(V_{\rm pg}^{(0)}) - \mu_N(V_{\rm pg}^{(0)})}{|e|\alpha_{\rm pg}}.$$

Coulomb blockade diamonds. The measurement of Coulomb blockade diamonds, as shown in Fig. 18.3, can be immediately understood on the basis of the above terminology and the empirical finding expressed by eq. (18.5). A diamond measurement corresponds to taking many plunger gate sweeps at various source–drain voltages. Applying a finite source–drain voltage $V_{\rm SD}$ to a quantum dot corresponds to opening a so-called *bias window*, which denotes the energetic region between $\mu_{\rm S}$ and $\mu_{\rm D}$, as shown as the light gray region in Fig. 18.11. The width of this energy

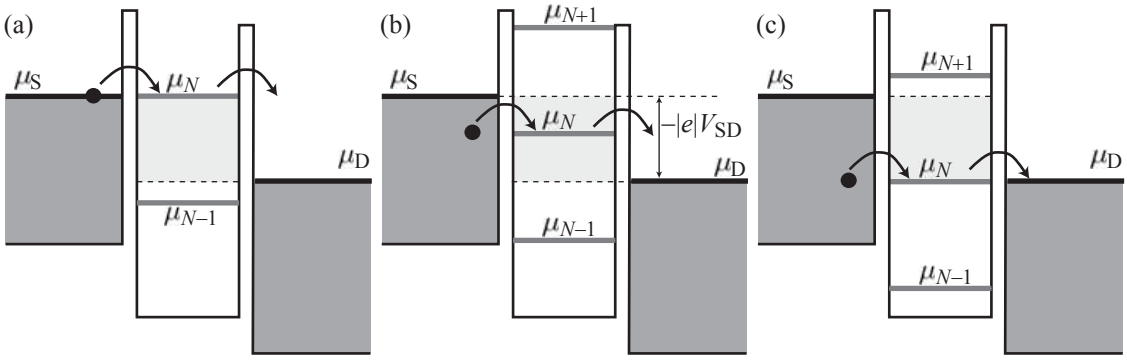


Fig. 18.11 Schematic representation of a quantum dot system with finite applied bias for various plunger gate voltages. The energy region in light gray represents the so-called bias window. Arrows indicate electron transfer. (a) Current onset for $\mu_S = \mu_N(V_{pg})$. (b) Situation with $\mu_S > \mu_N(V_{pg}) > \mu_D$ (region of current flow). (c) Current onset at $\mu_D = \mu_N(V_{pg})$.

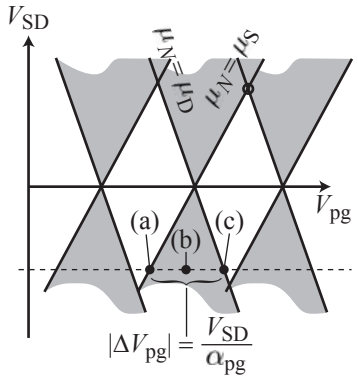


Fig. 18.12 Schematic Coulomb blockade diamonds. Current can flow in the light gray triangular-shaped regions. Black lines indicate alignment between one of the electrochemical potentials in the leads (μ_S or μ_D) with the electrochemical potential μ_N in the dot. The points labeled (a)–(c) refer to the situations represented schematically in Fig. 18.11. The encircled point marks the situation where $\mu_{N+1}(V_{SD}) = \mu_S$ and $\mu_N(V_{pg}) = \mu_D$.

window is given by $\mu_S - \mu_D = -|e|V_{SD}$. At finite V_{SD} a current can flow as a result of electron transfer between source and drain (see arrows in Fig. 18.11), as long as

$$\mu_S \geq \mu_N(V_{pg}) \geq \mu_D.$$

Increasing the plunger gate voltage, the current will rise from zero to a finite value when $\mu_S = \mu_N(V_{pg}^{(0)})$ [Fig. 18.11(a)]. The current will then remain until $\mu_D = \mu_N(V_{pg}^{(0)} + \Delta V_{pg}) = \mu_N(V_{pg}^{(0)}) - |e|\alpha_{pg}\Delta V_{pg}$, where it drops to zero again [see Fig. 18.11(c)]. The width in plunger gate voltage of the current-carrying situation is obtained from the difference of the two above equations to be

$$|\Delta V_{pg}| = \frac{|V_{SD}|}{\alpha_{pg}}. \tag{18.8}$$

In other words, there is a finite plunger gate voltage range allowing a current to flow, the size of which increases proportionally to the source–drain voltage. This equation defines triangular regions emerging from the zero source–drain voltage conductance resonances that extend to finite source–drain voltage as shown in Fig. 18.12. The constant of proportionality is the inverse lever arm α_{pg}^{-1}. As a consequence, the inverse lever arm can be read directly from Coulomb blockade diamond measurements (see Fig. 18.12).

As the triangles of current flow increase in ΔV_{pg} with increasing V_{SD}, corresponding triangles emerging from neighboring conductance resonances will intersect. Of particular interest are intersection points where $\mu_N(V_{pg}) = \mu_D$ and simultaneously $\mu_{N+1}(V_{SD}) = \mu_S$. At this point, which is encircled in Fig. 18.12, we have

$$-|e|V_{SD} = \mu_{N+1}(V_{pg}) - \mu_N(V_{pg}).$$

It therefore provides us with an energy calibration for the transport measurements. An estimate of this energy difference and an insight into

its meaning can be obtained from eqs (18.2) and (18.4) which lead to

$$\mu_{N+1}(V_{\rm pg}) - \mu_N(V_{\rm pg}) \approx \frac{e^2}{C} + \Delta,$$

where the first term represents the charging energy (which often dominates), and the second term represents the single-particle excitation energy.

The boundaries of the gray triangles in Fig. 18.12 (black lines) lead to peaks in measurements of the differential conductance $dI/dV_{\rm SD}$. The intersection points needed for the energy calibration can therefore be best determined from measurements of this quantity.

Excited state spectroscopy. So far we have been concerned only about elastic (i.e., energy conserving) electron transfer through the quantum dot involving quantum dot ground states, i.e., transitions $(N-1,0) \rightarrow (N,0) \rightarrow (N-1,0)$. In analogy to these ground state transitions we can also consider elastic transitions involving excited quantum dot states, such as $(N-1,0) \rightarrow (N,1) \rightarrow (N-1,0)$, or more generally $(N-1,m) \rightarrow (N,n) \rightarrow (N-1,m)$. The energy required for adding an electron to the quantum dot which is initially in the excited $N-1$ electron state m and ends up in the N electron state n is given by

$$\mu_N^{(n,m)}(V_{\rm pg}) = E_N^{(n)}(V_{\rm pg}) - E_{N-1}^{(m)}(V_{\rm pg}). \tag{18.9}$$

All $\mu_N^{(n,0)}$ with $n > 0$ are larger than the electrochemical potential $\mu_N \equiv \mu_N^{(0,0)}$, whereas all $\mu_N^{(0,m)}$ with $m > 0$ are smaller. As a consequence, each ground state transmission channel contributing to electron transport is surrounded by a bunch of excited state transmission channels as indicated in Fig. 18.13. With the help of eq. (18.5) and the definition of $\mu_N^{(n,m)}$ in eq. (18.9) we find for the gate voltage dependence of excited state transitions

$$\mu_N^{(n,m)}(V_{\rm pg}) = \mu_N^{(n,m)}(V_{\rm pg}^{(0)}) - |e|\alpha_{\rm pg}\Delta V_{\rm pg},$$

which shows that excited state transitions are shifted by the plunger gate parallel to ground state transitions [cf., eq. (18.7)].

The transmission channels involving excited states can contribute to electron transfer through the dot if they are within the bias window *and* if μ_N is in the bias window. The reason for the latter condition is the following: if μ_N is below $\mu_{\rm D}$ in Fig. 18.13, the dot will start any energy transfer from its lowest energy state $E_N^{(0)}$, and the $\mu_N^{(n,m)}$ are not relevant, but rather the $\mu_{N+1}^{(n,m)}$. On the other hand, if μ_N is above $\mu_{\rm S}$ in Fig. 18.13, the dot will start any electron transfer from its lowest energy state $E_{N-1}^{(0)}$ and therefore transitions starting from excited $N-1$ electron states are usually not relevant.

We now consider the consequences of the presence of excited states for Coulomb diamond measurements. For simplicity we assume, that only the excited transition $\mu_N^{(1,0)}$ is relevant. As the plunger gate voltage is

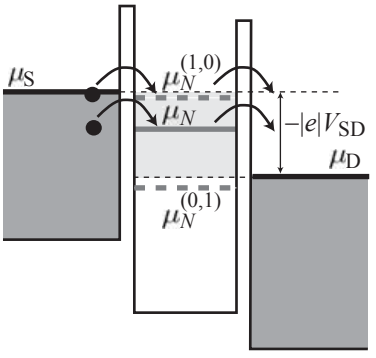


Fig. 18.13 Schematic representation of a quantum dot at finite source–drain voltage. In addition to the ground state transition μ_N, two excited state transitions are indicated as dashed lines. The upper one, $\mu_N^{(1,0)}$ is in the transport window and therefore contributes to electron transport through the dot.

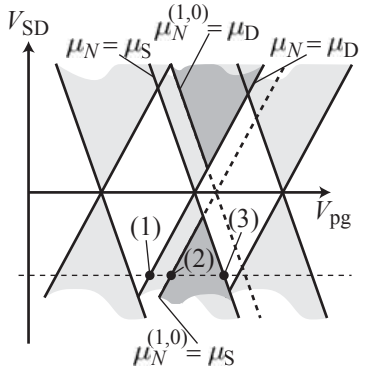


Fig. 18.14 Schematic Coulomb blockade diamonds with an excited state. Current can flow in the light gray triangular shaped regions. Black lines indicate alignment between one of the electrochemical potentials in the leads (μ_S or μ_D) with a transition $\mu_N^{(n,m)}$ in the dot. The points labeled (1)–(3) refer to the three current steps introduced in the text.

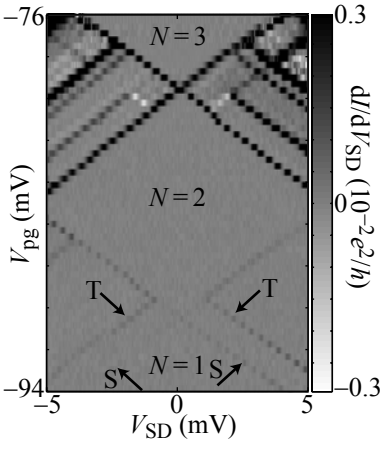


Fig. 18.15 Measured Coulomb blockade diamonds with excited states. The gray scale represents the differential conductance dI/dV_{SD}. Lines of states appear digitized as a result of the finite resolution in V_{pg} and V_{SD}. Note that the two axes are here interchanged compared to Fig. 18.14. Transitions labeled S and T involve the singlet ground state and the triplet excited states of the two-electron system (quantum dot helium).

increased, μ_N will approach μ_S, and a current step (step 1) will occur at $\mu_N = \mu_S$ (see the point labeled (1) in Fig. 18.14). Another step in the current is expected when $\mu_N^{(1,0)} = \mu_S$ (step 2). We will see later that this can be a step up or down depending on the tunneling coupling of the corresponding states to source and drain. A third step of the current, down to zero, will occur at even higher plunger gate voltage, where $\mu_N = \mu_D$ (step 3). The separation in plunger gate voltage of steps 1 and 3 are given by eq. (18.8). The separation in plunger gate voltage of steps 1 and 2 are given by

$$\Delta V_{pg} = \frac{\mu_N^{(1,0)}(V_{pg}^{(0)}) - \mu_N(V_{pg}^{(0)})}{|e|\alpha_{pg}} = \frac{E_N^{(1)}(V_{pg}^{(0)}) - E_N^{(0)}(V_{pg}^{(0)})}{|e|\alpha_{pg}}.$$

The energy difference in the numerator of the last expression is the excitation energy of the N-electron system. The simplest estimate would identify this excitation energy with the single-particle excitation energy Δ from eq. (18.4). Coulomb blockade measurements can therefore be used for excited state spectroscopy.

In measurements of the differential conductance, excited states appear as lines outside the diamonds where the current flow is Coulomb blockaded. This can be seen in Fig. 18.15 which shows a measurement of the differential conductance dI/dV_{SD} taken on a quantum dot fabricated by AFM lithography with a small electron number between $N = 1$ and 3. Lines representing electron transfers involving excited dot states can be seen outside the Coulomb-blockaded diamond shaped regions.

18.2 Quantum dot states

18.2.1 Overview

The more rigorous theoretical description of the Coulomb blockade effect can be split in two separate steps: In the first step the states of the isolated island are described using the general hamiltonian (8.1) of the closed system which neglects the tunneling coupling to source and drain. In the second step the tunneling coupling of quantum dot states is introduced as a small perturbation, and the current is calculated.

The exact diagonalization of the hamiltonian (8.1) is generally not possible if the quantum dot contains a large number (say, more than 10) electrons. In particular, the Coulomb interaction is responsible for the fact that the problem is hard to solve.

The hamiltonian H_N in (8.1) is the sum of the single-particle operators

$$h(\mathbf{r}) = -\frac{\hbar^2}{2m^\star}\Delta - e\int_V dV' \rho_{ion}(\mathbf{r}')G(\mathbf{r},\mathbf{r}') + \frac{e^2}{2}G(\mathbf{r},\mathbf{r}) - e\sum_i \phi_i\alpha_i(\mathbf{r})$$

and the two-particle operators

$$V(\mathbf{r},\mathbf{r}') = e^2 G(\mathbf{r},\mathbf{r}'),$$

and we can write

$$H_N = \sum_{n=1}^{N} h(\mathbf{r}_n) + \sum_{n=1}^{N} \sum_{m=1}^{n-1} V(\mathbf{r}_m, \mathbf{r}_n). \tag{18.10}$$

Different approximations for solving the many-particle problem have been used in the literature. Of particular importance for many-electron quantum dots are the capacitance model, the constant-interaction model, and the Hartree or the Hartree–Fock approximations which often lead to intuitive results because the system can be described with single-particle wave functions. States in quantum dots with few electrons (less than 10) can also be calculated using the configuration interaction method (sometimes also called 'exact diagonalization'). In the following we will give an overview of these methods starting with the simplest and most intuitive capacitance model, the results of which are identical to the constant interaction model.

18.2.2 Capacitance model

We describe the quantum dot as a metallic island with a discrete energy spectrum. In this description the interaction part of the hamiltonian (18.10), i.e., interaction effects of the electrons between each other and with the gate electrodes, are represented by a capacitance matrix. The charges on the individual metallic objects are then related to the electrostatic potentials by the equation

$$Q_i = \sum_{j=0}^{n} C_{ij} V_j + Q_i^{(0)}. \tag{18.11}$$

Here the indices $i = 1, 2, 3, \ldots, n$ denote the gate electrodes; the quantum dot island has the index $i = 0$. The charges $Q_i^{(0)}$ reside on the gates and the dot if all $V_i = 0$. The potential of the island is, however, unknown in general, but its charge is known to be an integer multiple of the elementary charge. We can therefore write

$$V_0(Q_0) = \frac{Q_0 - Q_0^{(0)}}{C_\Sigma} - \sum_{j=1}^{n} \frac{C_{0j}}{C_\Sigma} V_j,$$

where $C_\Sigma \equiv C_{00} = -\sum_{i=1}^{n} C_{0i} > 0$. The electrostatic energy needed to add N additional electrons to the quantum dot is given by

$$E_{\mathrm{elstat}}(N) = \int_{Q_0^{(0)}}^{Q_0^{(0)} - |e|N} dQ_0 V_0(Q_0) = \frac{e^2 N^2}{2C_\Sigma} + |e|N \sum_{j=1}^{n} \frac{C_{0j}}{C_\Sigma} V_j.$$

The remaining part of the hamiltonian (18.10) which has not been replaced by the capacitive description contains only single-particle operators and can therefore be seen as a quantum mechanical single-particle problem. If we assume that the solution of this single-particle problem

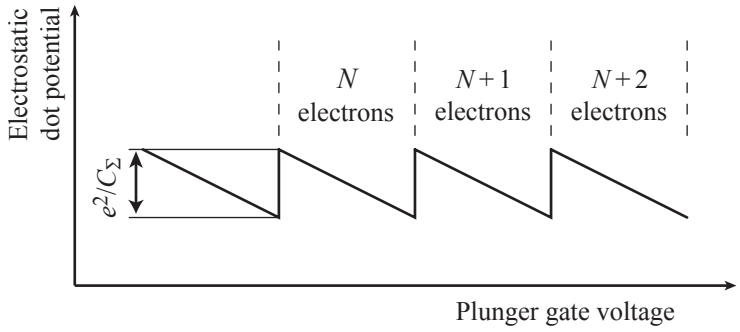


Fig. 18.16 Oscillations of the electrostatic potential in a quantum dot with gate voltage. The jumps are caused by individual electrons being added to the quantum dot.

gives energy levels $\epsilon_n^{(0)}$, the total energy of the island with N additional electrons is

$$E(N) = \sum_{n=1}^{N} \epsilon_n^{(0)} + \frac{e^2 N^2}{2C_\Sigma} + |e|N \sum_{i=1}^{n} \frac{C_{0i}}{C_\Sigma}(V_i - V_i^{(0)}). \tag{18.12}$$

Here we have included the term $|e|N \sum_{i=1}^{n} \frac{C_{0i}}{C_\Sigma} V_i^{(0)}$ in the quantization energy. In this model the electrochemical potential of the quantum dot is

$$\begin{aligned} \mu_N &= E(N) - E(N-1) \\ &= \epsilon_N^{(0)} + \frac{e^2}{C_\Sigma}\left(N - \frac{1}{2}\right) + |e| \sum_{i=1}^{n} \frac{C_{0j}}{C_\Sigma}(V_i - V_i^{(0)}). \end{aligned} \tag{18.13}$$

For large electron numbers N we have $N - 1/2 \approx N - 1$ and we will later see that we have exactly produced the expression (18.26) of the constant interaction model to be discussed later, provided that we identify the charging energy with $V_\mathrm{c} = e^2/C_\Sigma$ and the lever arm of gate i with $\alpha_i = -C_{0i}/C_\Sigma$.

The physical interpretation of the expression for the electrochemical potential of the quantum dot is the following: The energy $\epsilon_N^{(0)}$ is the chemical potential needed to add the Nth electron to the quantum dot. The remaining contribution

$$\frac{e^2}{C_\Sigma}\left(N - \frac{1}{2}\right) + |e| \sum_{i=1}^{n} \frac{C_{0j}}{C_\Sigma}(V_i - V_i^{(0)}),$$

consisting of the charging energy term and the gate-voltage-dependent term is the electrostatic potential. It shows a zigzag behavior as electrons are added to the island within increasing gate voltages as depicted in Fig. 18.16, because as long as μ_N is larger than the electrochemical potential in the source and drain contact (μ_s), N remains constant and the electrostatic potential changes linearly with the gate voltage. At values of the gate voltages where $\mu_N = \mu_\mathrm{s}$, the charge of the quantum jumps by one electronic charge and the electrostatic potential jumps upwards by e^2/C_Σ. This jump in the electrostatic potential has been

measured in the experiment shown in Fig. 18.4 where a quantum point contact detector was used to sense the charge in the quantum dot.

Separation of conductance peaks in plunger gate voltage. According to the qualitative model of the Coulomb blockade effect in Fig. 18.9(b) and (c) we expect (at negligibly small source–drain voltage) a conductance peak, if $\mu_{N+1} = \mu_S = \mu_D$. Within the constant interaction model we obtain from this relation the values of the plunger gate at which conductance peaks occur:

$$V_{pg}(N+1) = \frac{1}{e\alpha_{pg}} \left(\epsilon_{N+1} + V_c \cdot N - |e| {\sum_i}' \alpha_i V_i - \mu_S \right). \qquad (18.14)$$

The primed sum excludes summation over the plunger gate index. The separation of two neighboring conductance peaks in gate voltage is then

$$\Delta V_{pg} = V_{pg}(N+1) - V_{pg}(N) = \frac{1}{|e|\alpha_{pg}} \left(\epsilon_{N+1} - \epsilon_N + V_c \right),$$

i.e., the sum of the charging energy and the single-particle level separation.

Boundaries of Coulomb blockade diamonds. Using the capacitance model, the equations of the Coulomb blockade diamond boundaries can be determined for various biasing conditions. In the qualitative description of section 18.1.4 we have considered only the shift of the quantum dot electrochemical potential with plunger gate voltage at fixed source–drain voltage [eq. (18.7)]. Here we will confirm and complement these earlier results on the basis of the capacitance model of quantum dots. Using eq. (18.13), the capacitive action of the source and drain contact can be taken into account in addition to the plunger gate voltage. For calculating the Coulomb blockade diamond boundaries, we assume that the voltage $-fV_{SD}$ is applied to the drain and $(1-f)V_{SD}$ to source. The quantity f is a constant with $0 \le f \le 1$ which we introduce here for the following reason: if we choose $f = 0$, the drain contact is the reference potential of the circuit to which all applied voltages refer, and the full source–drain voltage is applied to the source. For $f = 1/2$, however, the voltage is antisymmetrically applied to source and drain, i.e., $-V_{SD}/2$ to drain and $+V_{SD}/2$ to source, and the reference point for all voltages is exactly between the two. These two situations are typically realized in experiments, but lead to slightly different results, as we will see below. We further determine in the experiment a plunger gate voltage $V_{pg}^{(0)}$ for which, at zero source–drain voltage, we define

$$\mu_N(V_{pg}^{(0)}) = \epsilon_N^{(0)} + \frac{e^2}{C_\Sigma} \left(N - \frac{1}{2} \right) - e\alpha_{pg} V_{pg}^{(0)} = \mu_{S/D} = 0.$$

In the following we use $\Delta V_{pg} = V_{pg} - V_{pg}^{(0)}$ and $\Delta_{N+1} = \epsilon_{N+1}^{(0)} - \epsilon_N^{(0)}$. The four Coulomb blockade diamond boundaries are given by the four linear

equations $\mu_N = \mu_S$, $\mu_N = \mu_D$, $\mu_{N+1} = \mu_S$, and $\mu_{N+1} = \mu_D$. These equations lead to the linear boundary equations

$$\Delta V_{pg} = \underbrace{-\frac{1}{\alpha_{pg}}\left[\alpha_S(1-f) - \alpha_D f + (1-f)\right]}_{m_1} V_{bias} \quad (18.15)$$

$$\Delta V_{pg} = \underbrace{-\frac{1}{\alpha_{pg}}\left[\alpha_S(1-f) - \alpha_D f - f\right]}_{m_2} V_{bias} \quad (18.16)$$

$$\Delta V_{pg} = \frac{1}{e\alpha_{pg}}\left(\Delta_{N+1} + \frac{e^2}{C_\Sigma}\right) - \frac{1}{\alpha_{pg}}\left[\alpha_S(1-f) - \alpha_D f + (1-f)\right] V_{bias} \quad (18.17)$$

$$\Delta V_{pg} = \frac{1}{e\alpha_{pg}}\left(\Delta_{N+1} + \frac{e^2}{C_\Sigma}\right) - \frac{1}{\alpha_{pg}}\left[\alpha_S(1-f) - \alpha_D f - f\right] V_{bias}. \quad (18.18)$$

Lever arms from diamond measurements. Equations (18.15) and (18.17), as well as (18.16) and (18.18) describe pairs of lines with the same slope but offset relative to each other by a certain amount. The magnitude of the difference of slopes, $\Delta m = m_1 - m_2$, gives

$$|\Delta m| = \frac{1}{\alpha_{pg}}, \quad (18.19)$$

i.e., the lever arm of the plunger gate, independent of the value of f.

The lever arms α_S and α_D can be determined from diamond measurements as well. If the source–drain voltage is applied such that $f = 0$ one finds

$$\frac{1}{\alpha_S} = \frac{m_1}{m_2} - 1 \text{ (drain grounded)},$$

and if it is applied such that $f = 1$ one obtains

$$\frac{1}{\alpha_D} = \frac{m_2}{m_1} - 1 \text{ (source grounded)}.$$

In the case of $f = 1/2$, only the difference

$$\alpha_S - \alpha_D = \frac{m_1 + m_2}{m_1 - m_2} \text{ (antisymmetric} V_{SD})$$

can be determined. In particular, if $m_1 = -m_2$, both lever arms are the same and the capacitive coupling of source and drain to the dot is the same.

Charging energy and single-particle level spacing from diamonds. Furthermore the sum of the charging energy and the single-particle level spacing can be read from diamonds. The two boundary lines (18.15) and (18.18) intersect at

$$V_{\text{bias}} = \frac{1}{e}\left(\Delta_{N+1} + \frac{e^2}{C_\Sigma}\right), \tag{18.20}$$

regardless of the value of f. In Fig. 18.3 the value of this sum is between 0.7 and 1 meV, depending on the specific diamond; in Fig. 18.15 it is beyond the boundaries of the plot at about 6 meV.

18.2.3 Approximations for the single-particle spectrum

The hamiltonian with the expectation value $\epsilon_n^{(0)}$ does not contain any electron–electron interaction. The electrons move in a confinement potential which is entirely determined by the gate-induced potentials and by the potentials of the fixed charges in the system. A popular approximation for the single-particle spectrum $\epsilon_n^{(0)}$ is the Fock–Darwin spectrum (Fock, 1928; Darwin, 1930) describing the motion of an electron in a two-dimensional isotropic harmonic oscillator subject to a magnetic field normal to the plane. The corresponding hamiltonian is

$$H = \frac{(\mathbf{p} + |e|\mathbf{A})^2}{2m^\star} + \frac{1}{2}m^\star\omega_0^2 r^2.$$

As a result of the cylindrical symmetry of the confinement potential, angular momentum is a good quantum number and states can be classified according to the number n of nodes of the wave function in radial direction and the angular momentum quantum number l. The resulting energy spectrum

$$E_{n,l} = \hbar\Omega\,(2n + |l| + 1) - \frac{1}{2}\hbar\omega_{\text{c}} l, \tag{18.21}$$

where $\Omega = (\omega_0^2 + (\omega_{\text{c}}/2)^2)^{1/2}$ and $\omega_{\text{c}} = eB/m^\star$ is shown in Fig. 18.17.

At zero magnetic field there is a ladder of states with increasing degeneracy, allowing us to define an s-shell ($l = 0$) taking two electrons of opposite spin, a p-shell ($l = \pm 1$) taking another four electrons, and so on. The numbers written in the gaps between these states indicate the total number of electrons that a system would have when all the shells below the number are completely filled. We will see later that shell filling has indeed been observed in small quantum dots with cylindrical symmetry. A finite magnetic field splits the states that are degenerate at zero field. In the limit of very large magnetic field, all states with $n = 0$ and $l > 0$ tend asymptotically towards the lowest Landau level of a two-dimensional system without confinement. The states with $n = 1$ and $l > 0$, and the state $n = 0, l = -1$, tend towards the second Landau level, and so on. Since the number of electrons in a quantum dot

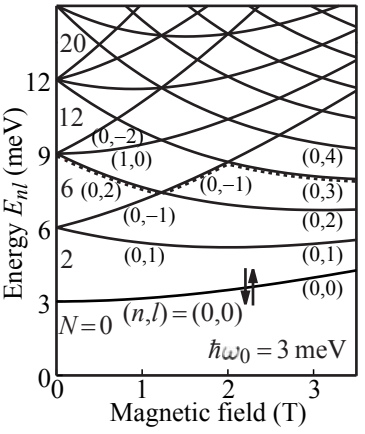


Fig. 18.17 Fock–Darwin spectrum calculated for a confinement energy $\hbar\omega_0 = 3$ meV. Each state can be occupied with spin up and spin down.

is fixed when the magnetic field is changed, the highest occupied state changes with magnetic field as levels cross (see dashed line in Fig. 18.17). This will lead to characteristic kinks in conductance peak positions as a function of magnetic field.

This model is close to reality only for small electron numbers (for which the Hartree–Fock approximation is often not very good), because there the confinement potential can be expanded around its minimum up to second order leading to the harmonic oscillator potential. At small electron numbers the deviations from the isotropic parabolic approximation are typically small and can be considered, for example, by perturbation theory. A more refined single-particle model for small electron numbers is the anisotropic harmonic oscillator in a magnetic field which can also be solved analytically (Schuh, 1985). For very large quantum dots, the electrons in the dot screen the confinement potential and a hard-wall confinement is in many cases superior to the parabolic approximations.

18.2.4 Energy level spectroscopy in a perpendicular magnetic field

Equation (18.14) gives the plunger gate voltages at which conductance resonances occur. The expression contains energy levels ϵ_N of the quantum dot. If measurements are made as a function of a magnetic field, conductance peaks shift in gate voltage. In many cases it is justified to assume that the charging energy e^2/C_Σ and the lever arms α_i are independent of the magnetic field. The shifts of conductance peak positions will then represent the magnetic field dependence of the single-particle energy levels, $\epsilon_N(B)$. Magnetotransport spectroscopy of energy levels is therefore a powerful tool to measure and identify single-particle energy levels in quantum dots.

Typically the energy spectra of quantum dots with many interacting electrons are complicated and hard to understand. The reason is that quantum dots, in contrast to atoms, usually do not have a highly symmetric confinement potential. As a result, statistical descriptions of energy levels have been used to describe statistical properties of conductance peak spacings and peak heights. They are known as *random matrix theory* (Beenakker, 1997; Alhassid, 2000; Aleiner *et al.*, 2002). A quantitative understanding of individual energy levels and their behavior in a magnetic field is only possible, if

(1) the number of electrons is very small (typically smaller than 10)
(2) the symmetry of the quantum dot is optimized.

Both conditions have been impressively fulfilled in so-called vertical quantum dots as they are depicted in Fig. 18.18. The material contains a $In_{0.05}Ga_{0.95}As$/AlGaAs quantum well with highly doped layers above and below the well. Vertical pillars with a diameter of about 500 nm are fabricated by etching. Subsequently, a gate electrode is evaporated at the base of the pillars which serves as the plunger gate. It has been

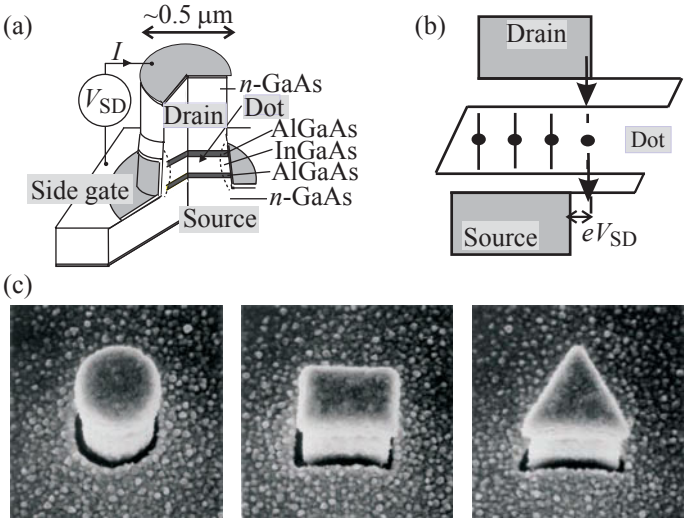


Fig. 18.18 Vertical quantum dots. (a) Cross section through a structure. (b) Energy level scheme. (c) Electron microscope images of quantum dots with different symmetry (Kouwenhoven *et al.*, 2001).

shown that in these structures the electron number can be increased from zero, one by one, to a finite number N. This is shown in Fig. 18.19. As a result of the rotational symmetry of the cylindrical quantum dots there is a shell structure similar to the one found in three-dimensional atoms. The confinement potential can approximately be described with the Fock–Darwin single-particle spectrum. These structures have therefore been called *artificial atoms*. The shell structure of the spectrum

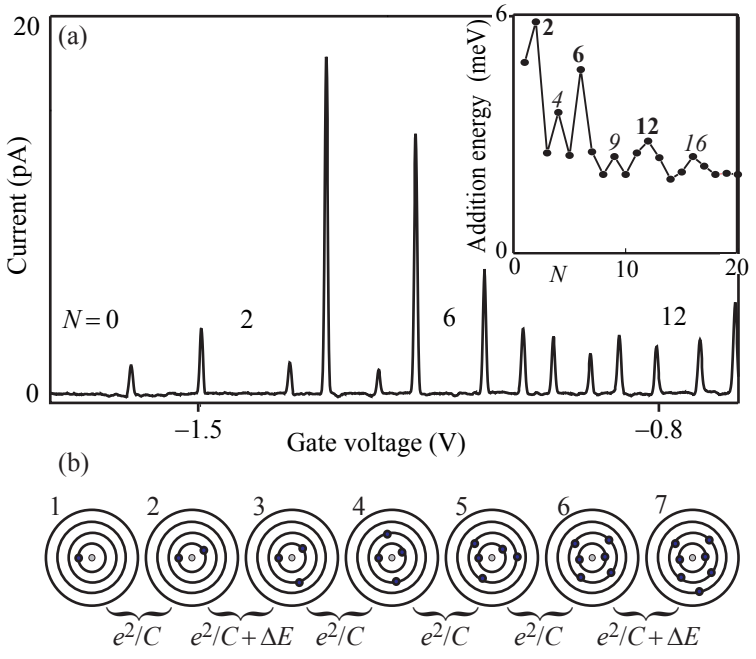


Fig. 18.19 (a) Coulomb blockade in artificial atoms. (b) Filling energy shells in analogy with three-dimensional atoms. (Kouwenhoven *et al.*, 2001).

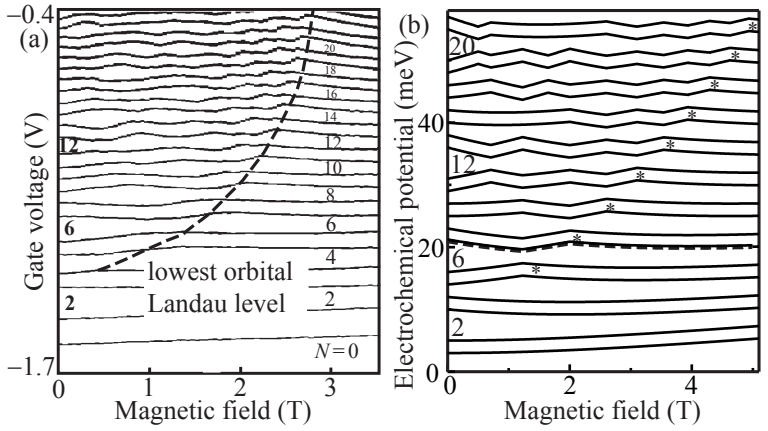


Fig. 18.20 (a) Measured magnetic field dispersion of conductance peaks in an artificial atom. (b) Fock–Darwin spectrum with a constant charging energy of 2 meV added between states of successive electron number (Kouwenhoven *et al.*, 2001).

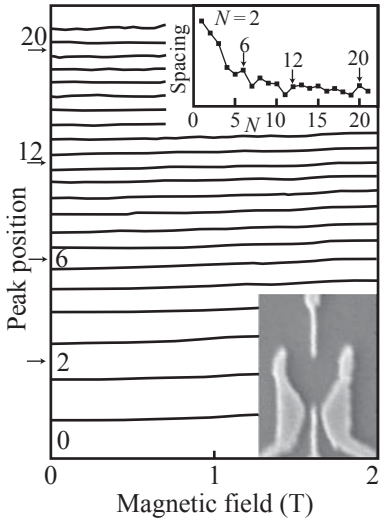


Fig. 18.21 Measured magnetic field dispersion of the conductance peaks in the few-electron lateral quantum dot depicted in the lower inset. The bottom inset shows the separation of neighboring conductance peaks vs. electron number showing the absence of clear shell filling. (Reprinted with permission from Ciorga *et al.*, 2000. Copyright 2000 by the American Physical Society.)

can be extracted from the separations between neighboring conductance peaks. When a second electron is filled into the lowest s-state, only the charging energy has to be supplied. With this electron, the s-shell is completely filled. The third electron has to be put into a level with p-symmetry, i.e., the charging energy and the single-particle level spacing between s- and p-shell has to be supplied. The p-level is two-fold degenerate due to spatial symmetry, and two-fold degenerate due to spin, i.e., there is a four-fold degeneracy. As a consequence, only the charging energy has to be supplied for further filling up to the sixth electron [see Fig. 18.19(b)]. Only charging the seventh electron costs an energy that is larger by the single-particle levels spacing to the d-level.

Figure 18.20(a) shows the measured magnetic field dependence of the conductance peak positions in this artificial atom. For comparison, the Fock–Darwin spectrum is shown in (b), where a constant charging energy of 2 meV has been added between successive states with increasing electron number. The agreement between the two spectra is not perfect, but this very simple model can reproduce many details of the measurement. Effects beyond the simple model occur because it neglects the exchange and correlation energies.

Quantum dots based on two-dimensional electron gases can be defined using lateral gates on the surface of a heterostructure. It turns out that lateral few-electron dots are not easy to fabricate. The reason is that in very small geometries the plunger gates have a significant electrostatic influence on the quantum point contacts coupling the dot to source and drain contacts. When the electron number is reduced, the tunneling contacts are in many geometries pinched off (current unmeasurably small), before the last quantized state has been depleted. However, in recent years an optimized gate geometry has been developed (Ciorga *et al.*, 2000) and extensively used, which allows us to define few-electron lateral dots. The structure, together with the magnetic field dependence of the conductance peak positions, is shown in Fig. 18.21. In contrast to the measurements performed on the highly symmetric vertical quantum dots, these spectra can be interpreted only using detailed model calcu-

lations that take the kidney-shaped confinement potential into account.

In special cases in can be possible to analyze spectra of lateral quantum dots with many electrons in detail without having to resort to statistical techniques. Here we discuss the particular case of a ring-shaped quantum dot which exhibits the cylindrical symmetry of the vertical quantum dots discussed before (Fuhrer *et al.*, 2001). We showed this ring structure in Fig. 6.15 where we discussed the AFM lithography technique. If the quantum point contacts connecting the ring to source and drain are in the tunneling regime, the ring is Coulomb blockaded. However, the transmission through the ring remains periodically modulated by an Aharonov–Bohm flux penetrating the ring. The reason is that the isolated ring exhibits an energy spectrum that is periodic in the flux quantum h/e. Figure 18.22(a) shows the measured energy spectrum of the ring as a function of the magnetic field. A constant charging energy of 190 µeV has been subtracted between neighboring conductance peaks. Most prominent are those states that perform a pronounced zigzag motion as a function of the magnetic field. This behavior reflects the motion of states in a perfect one-dimensional ring [see Fig. 18.22(b) and cf. eq. (14.5)]. As in artificial atoms, an angular momentum quantum number can be assigned to the different sections of the zigzag line. Different radial modes lead to families of zigzag states differing in slope (different angular momentum quantum numbers). States with a weak magnetic field dispersion arise because the source and drain contacts

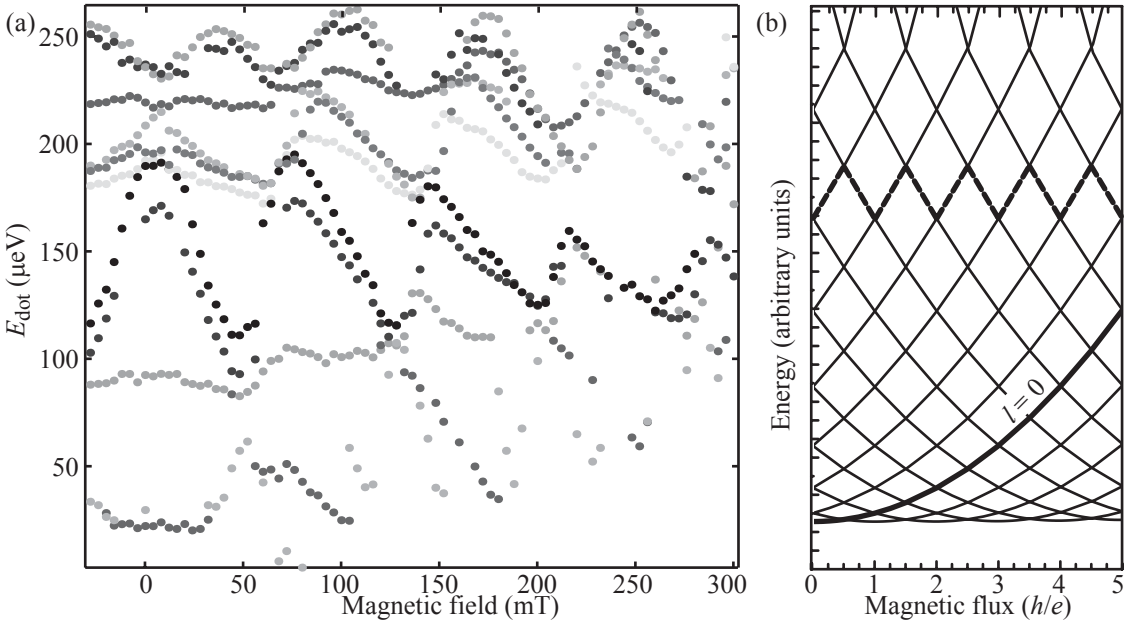


Fig. 18.22 (a) Measured magnetic field dispersion of the conductance resonances in a quantum ring. A constant charging energy of 190 µeV has been subtracted between neighboring states. (b) Calculated spectrum of an ideal one-dimensional quantum ring (Fuhrer *et al.*, 2001).

break the symmetry of the ring and lead to states of mixed positive and negative angular momentum.

18.2.5 Spectroscopy of states using gate-induced electric fields

In some cases the nature and location of single-particle states can be obtained by determining the lever arms of different gate electrodes on these states. Essentially geometric arguments decide on the behavior of the state: if, for example, a particular state is localized near a particular gate A, but further away from other gates, then the lever arm α_A will be significantly larger than all other lever arms. In Fig. 18.23 we again take the example of the quantum ring. It turns out that some states are localized in one of the two ring arms, whereas other states are extended around the whole ring [see Fig. 18.23(b)]. If a more positive gate voltage is applied on one side of the ring than on the other, as indicated by '+' and '−' in Fig. 18.23(a), the energy level of an extended state will be shifted very little, whereas the energy of a localized state is strongly shifted. In this way, the two classes of states can be distinguished through the different slopes in a plot of peak positions in the parameter plane of the two gate voltages, as shown in Fig. 18.23(c). The localized states exhibit a strong shift from the bottom left to the top right (dotted), whereas the extended states shift only weakly (solid).

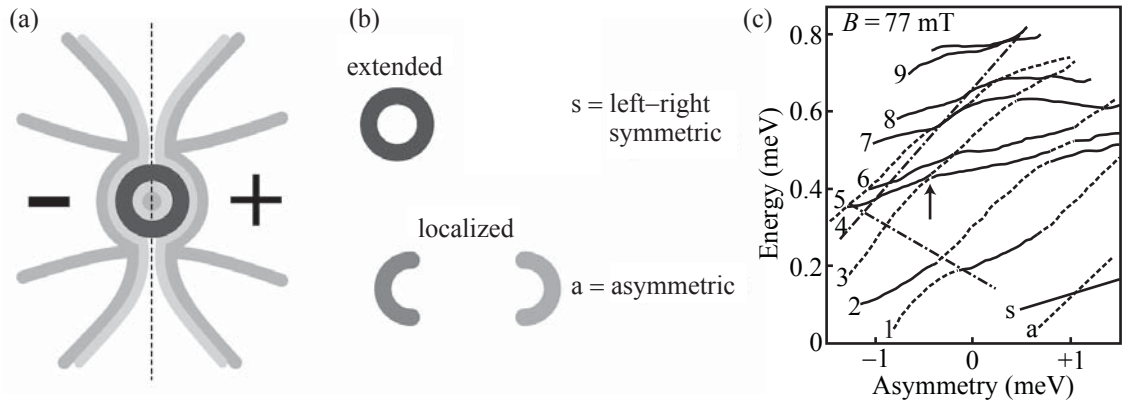


Fig. 18.23 (a) Schematic drawing of the quantum ring. (b) The extended and localized states of the ring. (c) Measured dependence of the conductance peak positions in a quantum ring on asymmetrically applied gate voltages. The asymmetry is the difference between the left and right gate voltages. A constant charging energy of 190 μeV between states has been subtracted. Flat states (solid lines) are extended around the ring, steep states (dashed lines) are localized in one arm of the ring. Dash-dotted lines indicate the positions of parametric charge rearrangements. (Reprinted with permission from Fuhrer *et al.*, 2003. Copyright 2003 by the American Physical Society.)

18.2.6 Spectroscopy of spin states in a parallel magnetic field

Spin states of lateral quantum dots based on two-dimensional electron gases can be identified using a parallel magnetic field. This orientation of the field is necessary because in a magnetic field perpendicular to the plane of the electron gas the strong orbital shifts of energy levels mask the much smaller shifts due to Zeeman splitting of levels. In a parallel magnetic field B the only orbital effect is a diamagnetic shift of energy levels which is proportional to B^2, and which is the same for all states. The linear magnetic field splitting due to the Zeeman effect is superimposed as a linear contribution with a sign that depends on the spin orientation of the tunneling electron. For single-particle levels we can therefore write

$$\epsilon(B_\parallel) = \gamma B_\parallel^2 + sg\mu_\mathrm{B} B_\parallel,$$

where the coefficient γ can be determined experimentally from the parabolic diamagnetic shift. Further, μ_B is Bohr's magneton, g is the Landé factor for electrons in the respective semiconductor material (e.g., $g = -0.44$ in GaAs) and $s = \pm 1/2$ is the spin quantum number along the direction of the magnetic field. Figure 18.24(b) shows an example for the Zeeman splitting of levels in the ring structure discussed earlier, now in an experiment in which the coefficient γ is very small and the Zeeman splitting can be directly seen in the raw data. The two neighboring conductance peaks arise from tunneling through the same orbital state. This state is first occupied with spin down and then with spin up. Figure 18.24(a) shows the dispersion of the same two conductance resonances in a perpendicular magnetic field on a much smaller field

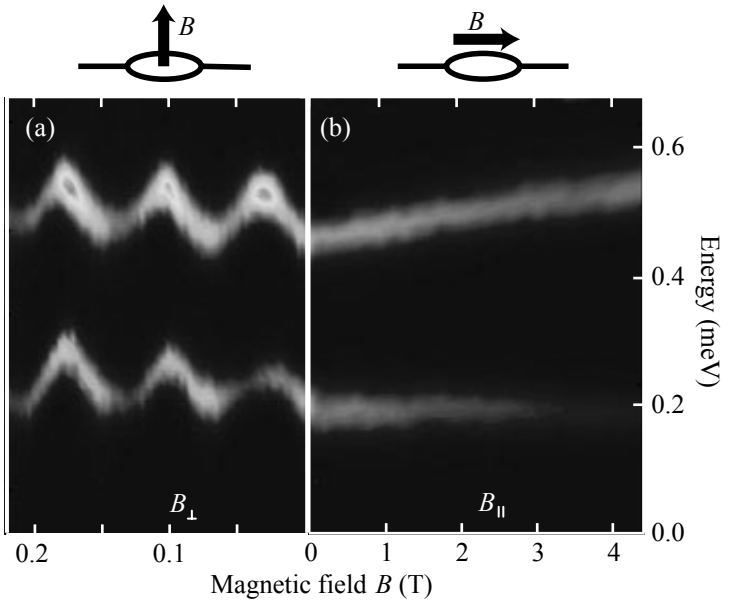


Fig. 18.24 (a) Dispersion of a spin pair in a perpendicular magnetic field. (b) Zeeman splitting of the same spin pair in a parallel magnetic field. (Ihn *et al.*, 2003. With kind permission of Springer Science and Business Media.)

scale. The parallel motion of both states indicates that they have the same orbital wave function.

18.2.7 Two electrons in a parabolic confinement: quantum dot helium

While the single-particle approximations together with the capacitive model for the interactions give a reasonable qualitative description of quantum dot physics, it is of great interest to study the interaction effects in more detail. In order to expose the problem we will discuss the simplest possible case here, which are two interacting electrons in a parabolic confinement potential. As the hamiltonian of the two-electron quantum dot we consider

$$H = \sum_{i=1}^{2} \left\{ \frac{[\mathbf{p}_i - e\mathbf{A}(\mathbf{r}_i)]^2}{2m^\star} + \frac{1}{2} m^\star \left[\omega_0^2 (x_i^2 + y_i^2) + \omega_z^2 z_i^2\right] \right\} + \frac{e^2}{4\pi\epsilon\epsilon_0 |\mathbf{r}_1 - \mathbf{r}_2|} + g^\star (\mathbf{s}_1 + \mathbf{s}_2)\mathbf{B}.$$

We assume the magnetic field to be normal to the plane of the electron gas, i.e., $\mathbf{B} = (0, 0, B)$, and choose the vector potential in the symmetric gauge $\mathbf{A} = B/2(-y, x, 0)$.

In general, two-electron problems are already quite involved as a result of the Coulomb interaction term. However, this particular hamiltonian separates into the energy of the center of mass motion and that of the relative motion if we introduce the center of mass and relative coordinates

$$\begin{aligned} \mathbf{R} &= \frac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2) \\ \mathbf{r} &= \mathbf{r}_1 - \mathbf{r}_2. \end{aligned}$$

We obtain a hamiltonian of the form (cf., Merkt *et al.*, 1991, Wagner *et al.*, 1992)

$$H = H_{\rm cm} + H_{\rm r} + H_{\rm Z}$$

where $H_{\rm cm}$ describes the center of mass motion, $H_{\rm r}$ the relative motion, and $H_{\rm Z}$ the Zeeman splitting. This separation of the hamiltonian is a peculiar property of the parabolic confinement potential. More complicated confinement potentials do not allow for such a separation. More specifically, the three parts of the total hamiltonian are

$$\begin{aligned} H_{\rm cm} &= \frac{[\mathbf{P} - e^\star \mathbf{A}(\mathbf{R})]^2}{2M^\star} + \frac{1}{2} M^\star \left[\omega_0^2 (X^2 + Y^2) + \omega_z^2 Z^2\right] \\ H_{\rm r} &= \frac{[\mathbf{p} - e'\mathbf{A}(\mathbf{r})]^2}{2\mu} + \frac{1}{2}\mu \left[\omega_0^2 (x^2 + y^2) + \omega_z^2 z^2\right] \\ &\quad + \frac{e^2}{4\pi\epsilon\epsilon_0 \sqrt{x^2 + y^2 + z^2}} \\ H_{\rm Z} &= g^\star (\mathbf{s}_1 + \mathbf{s}_2)\mathbf{B}. \end{aligned}$$

In the center of mass hamiltonian, the total mass $M^\star = 2m^\star$ and the total electronic charge $e^\star = 2e$ enters. The relative motion is governed by the reduced mass $\mu = m^\star/2$ and the reduced charge $e' = e/2$. Due to the separation of the hamiltonian into the three parts, we can write the total wave function of the system and the energy, respectively,

$$\begin{aligned} \Psi(\mathbf{r}_1, \mathbf{r}_2) &= \Phi(\mathbf{R})\psi(\mathbf{r})\chi(\mathbf{s}_1, \mathbf{s}_2) \\ E &= E_{\rm cm} + E_{\rm r} + E_{\rm Z}. \end{aligned}$$

This separation implies that excitations of the system can be classified to be either center of mass excitations, spin excitations, or excitations of the relative motion. The benefit of this separation is that we are now dealing with three separate problems which are much easier to solve. These three problems will be discussed in the following.

Center of mass motion. The center of mass hamiltonian represents a harmonic oscillator with different confinement in the plane and in the z-direction with a magnetic field applied in the z-direction. The solutions are the Fock–Darwin states and the Fock–Darwin energy spectrum (see Fig. 18.17) such that [cf., eq. (18.21)]

$$E_{\rm CM} = \hbar\Omega\,(2N + |M| + 1) + \frac{\hbar\omega_{\rm c}}{2}M + \hbar\omega_z\left(N_z + \frac{1}{2}\right),$$

where

$$\Omega := \sqrt{\omega_0^2 + \left(\frac{\omega_{\rm c}}{2}\right)^2}, \tag{18.22}$$

and M is the angular momentum quantum number for the center of mass motion, N is its radial quantum number, and N_z is the quantum number for motion in the z-direction. The frequency $\omega_{\rm c} = eB/m^\star = e^\star B/M$ is the cyclotron frequency.

The wave function of the center of mass motion is a product of the z-dependent wave function Φ_z and an in-plane wave function $\Phi_{\text{in-plane}}$. The latter is given by

$$\Phi_{\text{in-plane}}(\mathbf{R}) = \frac{1}{L_0}\sqrt{\frac{N!}{\pi(N+|M|)!}}e^{iM\varphi}\left(\frac{R}{L_0}\right)^{|M|} L_N^{|M|}(R^2/L_0^2)e^{-R^2/L_0^2}$$

with R being the length of the in-plane component of the vector $\mathbf{R}$, the generalized Laguerre polynomials L_n^ν (Abramowitz and Stegun, 1984) and the length scale

$$L_0 = \sqrt{\frac{2\hbar}{M\Omega}}.$$

The part of the wave function describing the center of mass motion in the z-direction is the eigenfunction of the harmonic oscillator.

Zeeman hamiltonian. The spin hamiltonian $H_{\rm Z}$ has the familiar singlet

$$|\,S\rangle = \frac{1}{\sqrt{2}}\left(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle\right)$$

and triplet states

$$\begin{aligned} |T_{-1}\rangle &= |\downarrow\downarrow\rangle \\ |T_0\rangle &= \frac{1}{\sqrt{2}}\left(|\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle\right) \\ |T_1\rangle &= |\uparrow\uparrow\rangle \end{aligned}$$

as solutions. In the magnetic field the triplet states split into three branches with spacing $g^\star \mu_\mathrm{B} S_z B$, where $S_z = 0, \pm 1$. The energies are

$$E_\mathrm{Z} = g^\star \mu_\mathrm{B} S_z B.$$

Since the total two-particle wave function must be antisymmetric with respect to a particle exchange and the center of mass wave function is symmetric, the antisymmetric spin singlet state will be only compatible with antisymmetric wave function of the relative motion while the symmetric triplet sate wave functions require an antisymmetric wave function of the relative motion.

Relative motion. After having found analytic solutions for the center of mass motion and the spin dynamics, the remaining challenge is the hamiltonian of the relative motion which contains the Coulomb repulsion between electrons. Owing to the axial symmetry of the problem this hamiltonian can conveniently be expressed in cylinder coordinates:

$$\begin{aligned} H_\mathrm{r} &= -\frac{\hbar^2}{2\mu}\Delta + \frac{\hbar e' B}{2\mu i}\frac{\partial}{\partial\varphi} + \frac{e'^2 B^2}{8\mu}\rho^2 + \frac{1}{2}\mu\left[\omega_0^2\rho^2 + \omega_z^2 z^2\right] \\ &\quad + \frac{e^2}{4\pi\epsilon\epsilon_0\sqrt{\rho^2+z^2}} \\ &= -\frac{\hbar^2}{2\mu}\left[\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\right) + \frac{\partial^2}{\partial z^2}\right] + \frac{e'^2 B^2}{8\mu}\rho^2 + \frac{1}{2}\mu\left[\omega_0^2\rho^2 + \omega_z^2 z^2\right] \\ &\quad -\frac{\hbar^2}{2\mu}\frac{1}{\rho^2}\frac{\partial^2}{\partial\varphi^2} + \frac{\hbar e' B}{2\mu i}\frac{\partial}{\partial\varphi} + \frac{e^2}{4\pi\epsilon\epsilon_0\sqrt{\rho^2+z^2}}. \end{aligned}$$

As a consequence of the axial symmetry, the angle dependence is described by the eigenfunctions of the z-components of angular momentum with quantum number m, i.e.,

$$\psi(\mathbf{r}) = \frac{1}{\sqrt{2\pi}} e^{im\varphi} u_m(\rho, z). \tag{18.23}$$

At this point we can state that due to the requirement that the total wave function $\Psi(\mathbf{r}_1, s_1; \mathbf{r}_2, s_2)$ must be antisymmetric when the two particles are interchanged (particle interchange means $\varphi \to \varphi + \pi$), states with even relative angular momentum m are spin singlet states, whereas states with odd m are spin triplet states.

Inserting this wave function, the hamiltonian for finding the function

$u(\rho, z)$ is then

$$\begin{aligned} H_\mathrm{r} &= -\frac{\hbar^2}{2\mu}\left[\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\right)+\frac{\partial^2}{\partial z^2}\right]+\frac{1}{2}\mu\left[\Omega^2\rho^2+\omega_z^2 z^2\right] \\ &\quad +\frac{\hbar^2}{2\mu}\frac{m^2}{\rho^2}+\frac{\hbar\omega_\mathrm{c} m}{2}+\frac{e^2}{4\pi\epsilon\epsilon_0\sqrt{\rho^2+z^2}}. \end{aligned}$$

Here we have introduced the cyclotron frequency $\omega_\mathrm{c} = eB/m^\star = e'B/\mu$ which replaces the magnetic field strength B, and the effective confinement frequency $\Omega = \sqrt{\omega_0^2+\omega_\mathrm{c}^2/4}$ already introduced in eq. (18.22).

The problem of the relative motion has now been simplified to a two-dimensional partial differential equation, where the radial and vertical coordinates are still coupled via the three-dimensional Coulomb interaction. Although we will not be able to find an analytic solution for this problem, further insight is gained by looking for characteristic length and energy scales of the system. Therefore we now introduce dimensionless coordinates and energies. The characteristic unit of length is the extent of the ground state wave function of a harmonic oscillator of mass $m^\star$ with confinement frequency Ω. This length is given by

$$l_0 = \sqrt{\frac{\hbar}{m^\star\Omega}}.$$

The new relative coordinates are therefore $\rho' = \rho/l_0$ and $z' = z/l_0$. Inserting them in the hamiltonian (and immediately omitting the primes) gives

$$\begin{aligned} H_\mathrm{r} &= \hbar\Omega\left\{-\left[\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\right)+\frac{\partial^2}{\partial z^2}\right]+\frac{1}{4}\left[\rho^2+\frac{\omega_z^2}{\Omega^2}z^2\right]+\frac{m^2}{\rho^2}\right\} \\ &\quad +\frac{\hbar\omega_\mathrm{c} m}{2}+\frac{e^2}{4\pi\epsilon\epsilon_0 l_0\sqrt{\rho^2+z^2}}. \end{aligned}$$

We now normalize the energy of the radial motion using the characteristic energy scale $\hbar\Omega$ and define

$$\epsilon_\mathrm{r} = \frac{1}{\hbar\Omega}\left(E_\mathrm{r}-\frac{1}{2}\hbar\omega_\mathrm{c} m\right)$$

and the dimensionless interaction parameter

$$\lambda := \frac{e^2/4\pi\epsilon\epsilon_0 l_0}{\hbar\Omega} = \sqrt{\frac{2E^\star_\mathrm{Ry}}{\hbar\Omega}}$$

which is the ratio of the Coulomb energy and the effective harmonic confinement energy. At zero magnetic field this parameter is sometimes called the Wigner parameter. The parameter λ is proportional to l_0 and therefore decreases with $1/\sqrt{B}$ at large magnetic fields, i.e., when $\omega_\mathrm{c} \gg \omega_0$. Figure 18.25(a) shows the magnetic field dependence of λ for GaAs, assuming values of $\hbar\omega_0$ between 1 meV and 5 meV in steps of 1 meV.

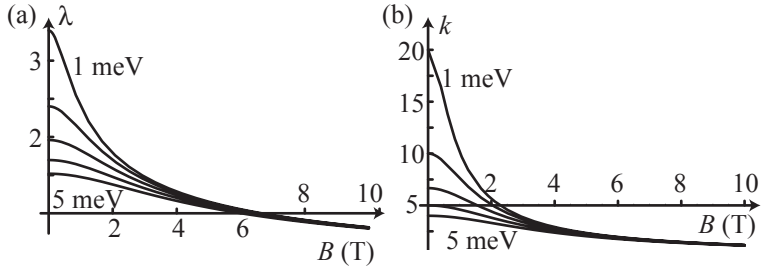


Fig. 18.25 (a) Values of the parameter λ as a function of magnetic field B for GaAs. The confinement energies $\hbar\omega_0$ are between 1 meV and 5 meV in steps of 1 meV. (b) Values of the parameter k as a function of magnetic field B for GaAs. The confinement energies $\hbar\omega_0$ are between 1 meV and 5 meV in steps of 1 meV. The confinement in the z-direction is $\hbar\omega_z = 20$ meV.

In addition, we define the parameter describing the three-dimensionality of the problem via

$$k := \frac{\omega_z}{\Omega}.$$

The geometry parameter k decreases with magnetic field because at $\omega_c \gg \omega_0$ the frequency $\Omega \propto B$ and therefore $k \propto 1/B$ [see Fig. 18.25(b)]. This means that the problem becomes increasingly three-dimensional in character as the magnetic field increases.

With these new parameters the eigenvalue problem becomes

$$\left\{ -\left[\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\right) + \frac{\partial^2}{\partial z^2}\right] + \frac{1}{4}\left[\rho^2 + k^2 z^2\right] + \frac{m^2}{\rho^2} + \frac{\lambda}{\sqrt{\rho^2+z^2}} \right\} u(\rho,z) = \epsilon_r u(\rho,z). \quad (18.24)$$

We can see that the motions in ρ and z are not separable due to the Coulomb repulsion term. The whole problem depends on the two parameters λ (interaction) and k (geometry).

The motion in the z-direction in eq. 18.24 can be taken into account approximately by replacing the pure Coulomb interaction with a Coulomb interaction averaged over the z-motion, assuming that $\omega_z \gg \omega_0$ (strong confinement in the z-direction). The resulting equation for the radial motion is (cf., Nazmitdinov *et al.* 2002) using our dimensionless quantities

$$\left\{ -\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\right) + \frac{1}{4}\rho^2 + \frac{m^2}{\rho^2} + \frac{2}{\pi}\frac{\lambda}{\sqrt{\rho^2+\Delta z_0^2}} K\left(\frac{\Delta z_0^2}{\rho^2+\Delta z_0^2}\right) \right\} u(\rho) = \tilde{\epsilon}_r u(\rho),$$

where $K(x)$ is the complete elliptic integral of the first kind (cf., Abramowitz and Stegun, 1984), $\Delta z_0^2 = 4 j_z \Omega/\omega_z$, $j_z = n_z + 1/2$ is the semiclassically quantized action variable for motion in the z-direction, and

$$\tilde{\epsilon}_r = \epsilon_r - \frac{\omega_z}{\Omega} j_z.$$

The second term represents the energy of harmonic oscillatory motion in the z-direction. For the lowest energy state we have $n_z = 0$ and $j_z = 1/2$.

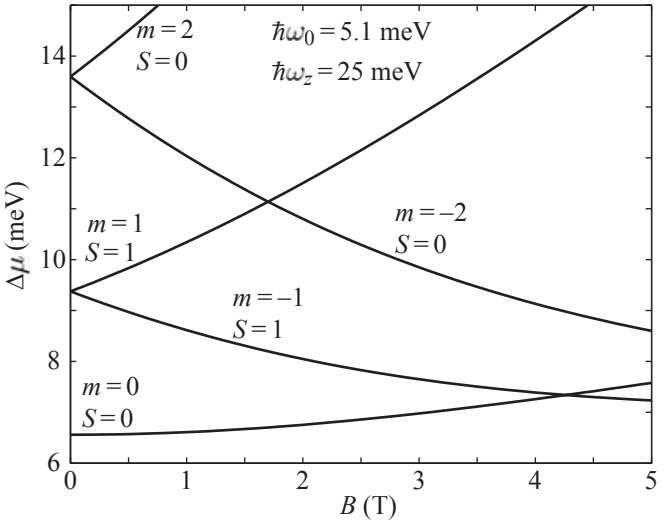


Fig. 18.26 Numerically calculated addition spectrum for quantum dot helium with a parabolic confinement potential. Only excitations of the relative motion are taken into account here. For the center of mass motion and the relative motion in the z-direction the ground state is assumed.

This approximate form of the problem demonstrates the way in which the effective thickness of the quantum dot Δz_0 enters the modified Coulomb interaction. For $\omega_z \to \infty$ we have $\Delta z_0^2 \to 0$ and, because $K(0) = \pi/2$, we recover the bare Coulomb interaction λ/ρ. The finite thickness essentially reduces the Coulomb interaction strength for electron–electron separations below Δz_0 and leaves it unchanged for larger distances. The interaction still diverges as $\rho \to 0$.

The remaining problem depends only on the radial coordinate ρ of the relative motion. The problem can therefore be solved numerically with little effort. Figure 18.26 shows the addition spectrum that has been calculated for the confinement parameters given in the figure, neglecting the Zeeman splitting. It can be seen that the ground state of quantum dot helium is a spin singlet ($S = 0$) with relative angular momentum $m = 0$. Increasing magnetic field forces the two electrons to move closer together thereby increasing the ground state energy. At a magnetic field of about 4.3 T, a singlet–triplet crossing occurs and the triplet state ($S = 1, m = -1$) becomes the ground state for higher magnetic fields. At zero magnetic field there is a degeneracy of the first excited triplet states due to the cylindrical symmetry of the problem. States with $m = \pm 1$ and $S_z = 0, \pm 1$ have the same energy. The orbital degeneracy may be lifted in systems where the confinement potential in the plane deviates from the isotropic shape assumed here. In the Coulomb diamonds of Fig. 18.15 tunneling through the singlet ground state (labeled S) and the triplet excited state (labeled T) is observed. In the symmetric case considered here, the finite relative angular momentum of the two electrons leads to net magnetic moments with opposite signs for the two states with $m = \pm 1$. This moment is responsible for the splitting of the degenerate states with magnetic field. The state with its magnetic moment parallel to the field ($m = -1$) will lower its energy, whereas the other state ($m = +1$) increases in energy. However, the spin degeneracy for the

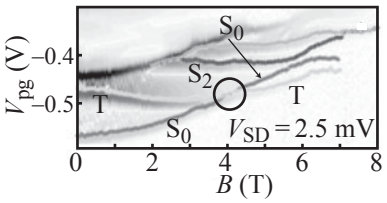


Fig. 18.27 Measured excitation spectrum of quantum dot helium in a magnetic field. The plotted quantity is $dI/dV_{\rm pg}$. The data was taken at finite source–drain voltage $V_{\rm SD} = 2.5\,\mathrm{mV}$. The lowest singlet state with relative angular momentum $m = 0$ is labeled 'S_0', the triplet excited state is 'T', and the next higher singlet excited state with $m = -2$ is 'S_2'. The singlet–triplet splitting occurring around 4 T is encircled. (Reprinted with permission from Ellenberger *et al.*, 2006. Copyright 2006 by the American Physical Society.)

triplet state remains, if we neglect the Zeeman interaction.

Indeed the measured excited state spectrum of quantum dot helium in a magnetic field including the singlet–triplet transition of the ground state can be brought into very good agreement with the calculated energy spectrum. Figure 18.27 shows such a measured spectrum exhibiting the singlet ground state with its energy increase in a magnetic field, the triplet excited states and the singlet–triplet crossing at finite magnetic field (encircled). In addition, a higher excited singlet state with angular momentum $m = -2$ can be seen.

18.2.8 Hartree and Hartree–Fock approximations

After having seen that even the two-electron problem can be rather tedious to solve, even in the simplest case of a harmonic confinement potential, we are now ready to appreciate the value of various very powerful approximations that lead to single-particle wave functions in the case of many interacting electrons.

Hartree approximation. Within the Hartree approximation the wave functions of the N-electron system are written as a product of (initially unknown) orthonormalized single-particle wave functions

$$\psi_N = \prod_{n=1}^{N} \varphi_n(\mathbf{r}_n, \{\phi_i\}).$$

The $\varphi_n(\mathbf{r}_n, \{\phi_i\})$ have to be determined self-consistently from the Hartree equation (8.12). The total energy of the N-electron system is then given by

$$E_{\rm H} = \sum_{n=1}^{N} \epsilon_n + \sum_{n=1}^{N} \sum_{m=1}^{n-1} C_{mn},$$

where

$$\epsilon_n \equiv \langle n\,|h(\mathbf{r}_n)|\,n\rangle = \int d^3r \varphi_n^\star(\mathbf{r}, \{\phi_i\}) h(\mathbf{r}) \varphi_n(\mathbf{r}, \{\phi_i\})$$

and

$$C_{mn} \equiv \langle mn\,|V(\mathbf{r}_m, \mathbf{r}_n)|\,mn\rangle = \int d^3r \int d^3r' \left|\varphi_m(\mathbf{r}, \{\phi_i\})\right|^2 V(\mathbf{r}, \mathbf{r}') \left|\varphi_n(\mathbf{r}', \{\phi_i\})\right|^2.$$

Before we interpret this result, we consider the same problem in the Hartree–Fock approximation.

Hartree–Fock approximation. The Hartree–Fock approximation takes into account that wave functions of fermions have to change sign if two arbitrary particles are exchanged. This symmetry requirement is

fulfilled by taking the total wave function to be a Slater determinant of single-particle wave functions, i.e.,

$$\psi_N = \frac{1}{\sqrt{N!}} \begin{vmatrix} \varphi_1(x_1, \{\phi_i\}) & \cdots & \varphi_N(x_1, \{\phi_i\}) \\ \vdots & & \vdots \\ \varphi_1(x_N, \{\phi_i\}) & \cdots & \varphi_N(x_N, \{\phi_i\}) \end{vmatrix}.$$

Here the coordinates x_i represent the spatial coordinate $\mathbf{r}_i$ and the spin coordinate s_i (z-component of the electron spin). The single-particle wave functions have to be determined self-consistently from the Hartree–Fock equation. Assuming that this has been done, we obtain for the total energy of the N-electron system

$$E_{HF} = \sum_{n=1}^{N} \epsilon_n + \sum_{n=1}^{N} \sum_{m=1}^{n-1} C_{mn} - \sum_{n=1}^{N} \sum_{m=1}^{n-1} X_{mn}.$$

In comparison to the Hartree approximation there is an additional interaction term called the exchange interaction. The corresponding matrix elements are

$$X_{mn} = \langle mn \,|V(\mathbf{r}_m, \mathbf{r}_n)|\, nm\rangle = \delta_{s_n s_m} \times$$
$$\int d^3r \int d^3r' \varphi_m^\star(\mathbf{r}, \{\phi_i\}) \varphi_n^\star(\mathbf{r}', \{\phi_i\}) V(\mathbf{r}, \mathbf{r}') \varphi_m(\mathbf{r}', \{\phi_i\}) \varphi_n(\mathbf{r}, \{\phi_i\}).$$

Their values depend on the spins of the two states n and m.

The Hartree–Fock approximation does not always deliver solutions that are eigenvectors of the total spin $\mathbf{S}^2$ of the system, but they are always eigenstates of the z-component of the total spin $S_z = \sum_i s_i$. For example, if we consider a system with three electrons, we would expect states with the six possible spin configurations $S = 1/2, S_z = \pm 1/2$, and $S = 3/2, S_z = \pm 3/2, \pm 1/2$. The Hartree–Fock approximation allows only four, namely, $S_z = \pm 1/2, \pm 3/2$. For the states with $S_z = \pm 3/2$ also the total spin $S = 3/2$ is a good quantum number. However, for the states with $S_z = \pm 1/2$ it is not a good quantum number because the total spin could be either $S = 1/2$ or $S = 3/2$. In the case of even electron number, the Hartree–Fock state in which all the lowest energy states are occupied with spin pairs is an eigenstate with $S = 0$ and $S_z = 0$.

Koopman's theorem. We now ask what energy is gained if a single electron is removed from the N-electron system. This could, for example, happen as the result of a tunneling process into the source or drain contact. In order to find this energy, we would have to solve the self-consistent problem for the N- and the $N-1$-electron system, and then determine the difference of their energies. *Koopman's theorem* states that, for large numbers N, we can safely make the approximation that the removal of the Nth electron does not alter the other wave functions $\varphi_n(\mathbf{r}, \{\phi_i\})$ for $n \neq N$. In essence this implies that we can neglect the

self-consistent electron rearrangement that may take place. With this approximation the energy gain is given by

$$\mu_N = \epsilon_N + \sum_{m=1}^{N-1} C_{mN} - \sum_{m=1}^{N-1} X_{mN}.$$

At the same time this is the energy that we require in order to add the Nth electron to the $N-1$-electron system. We therefore call μ_N the electrochemical potential of the quantum dot with $N-1$ electrons. The first term in μ_N depends on the electrostatic potentials of the gate electrodes, i.e., on the applied gate voltages. The second term is the so-called charging energy of the quantum dot (C being the abbreviation for charging energy) which one can already obtain from the Hartree approximation. We write the charging energy

$$V_H(N) = \sum_{m=1}^{N-1} C_{mN}.$$

The third term describes the exchange interaction (X stands for exchange energy). We can write it as

$$V_{xc}^{\uparrow/\downarrow}(N) = \sum_{m=1}^{N-1} X_{mN},$$

depending on the spin of the Nth electron. The contribution of the exchange energy is always negative. It creates a tendency towards spin alignment.

Gate lever arms. The single-particle energies ϵ_n are in the Hartree–Fock approximation determined for given electrostatic potentials ϕ_i of the gate electrodes. We can now ask how much the energies ϵ_n change, if the gate voltages are changed by small amounts. If we denote with $\epsilon_n^{(0)}$ the single-particle energies for the particular values $\phi_i^{(0)}$ of gate potentials, then

$$\epsilon_n^{(0)} = \left\langle n \left| -\frac{\hbar^2}{2m^\star}\Delta - e\int_V dV' \rho_{ion}(\mathbf{r}')G(\mathbf{r},\mathbf{r}') \right.\right. \\ \left.\left. + \frac{e^2}{2}G(\mathbf{r},\mathbf{r}) - e\sum_i \phi_i^{(0)}\alpha_i(\mathbf{r}) \right| n \right\rangle,$$

and we can write

$$\epsilon_n = \epsilon_n^{(0)} - e\sum_i (\phi_i - \phi_i^{(0)}) \left\langle n \left| \alpha_i(\mathbf{r}) \right| n \right\rangle.$$

The wave functions will not change appreciably for small changes of the gate voltages, and the energies ϵ_n shift linearly with the gate voltages. The quantity $\langle n |\alpha_i(\mathbf{r})| n\rangle$ is called the *lever arm* of gate i acting on state n. If we insert the definition of the lever arms in the equation for the electrochemical potential of the quantum dot, we obtain

$$\mu_N = \epsilon_N^{(0)} - e\sum_i (\phi_i - \phi_i^{(0)}) \left\langle n \left| \alpha_i(\mathbf{r}) \right| n \right\rangle + V_H(N) - V_{xc}^{\uparrow/\downarrow}(N). \quad (18.25)$$

What is the spin of the Nth electron? The fact that the matrix elements of the exchange interaction X_{mn} depend on the spins of the states m and n is of crucial importance for the total spin of the quantum dot ground state. We ask now whether it is energetically favorable to fill an additional electron with spin down ($\downarrow$) or spin up ($\uparrow$). Let us assume, we fill a $\downarrow$-electron into the level ϵ_k. The required energy is

$$\mu_N^{\downarrow} = \epsilon_k + \sum_{m\,\mathrm{occ.}} C_{mk} - \sum_{m\,\mathrm{occ.}} X_{m,k\downarrow}.$$

Correspondingly, the energy for filling an $\uparrow$-electron into the level ϵ_n is

$$\mu_N^{\uparrow} = \epsilon_n + \sum_{m\,\mathrm{occ.}} C_{mn} - \sum_{m\,\mathrm{occ.}} X_{m,n\uparrow}.$$

In the case $\mu_N^{\uparrow} < \mu_N^{\downarrow}$ the Nth electron will occupy the $\uparrow$-state, in the other case the $\downarrow$-state. An interesting situation occurs if the levels ϵ_k and ϵ_n are energetically degenerate and the charging energies $\sum_{m\,\mathrm{occ.}} C_{mk}$ and $\sum_{m\,\mathrm{occ.}} C_{mn}$ are also the same. In this case the difference of the exchange interactions

$$\xi = \sum_{m\,\mathrm{occ.}} X_{m,k\downarrow} - \sum_{m\,\mathrm{occ.}} X_{m,n\uparrow}$$

decides about the spin of the Nth electron.

In atoms this situation is, for example, known as partial p-shell filling. If there is already an electron in the p_x-orbital, the next electron can be filled either in the p_y- or the p_z-orbital with either spin $\uparrow$ or $\downarrow$. The single-particle energies and the charging energies are the same for both alternatives. However, the exchange interaction gives preference to parallel spins in the orbitals p_x and p_y (or p_z). The exchange interaction leads in this way to one of Hund's rules according to which degenerate levels will be filled first with single electrons (as a result of the Hartree energy) and parallel spins (as a result of the exchange interaction).

18.2.9 Constant interaction model

The already relatively complex self-consistent description of quantum dots can be further simplified if the following assumptions are made:

(1) The exchange interaction is small and can be neglected.

(2) The Hartree energy increases monotonously with N and we can write

$$V_H(N) = (N-1) \cdot \underbrace{\frac{1}{N-1} \sum_{m=1}^{N-1} C_{mN}}_{c(N)}.$$

The charging energy per electron, $c(N)$, will fluctuate in quantum dots with large electron number (typically $N > 100$) with increasing N around a constant average V_c, such that $c(N) = V_\mathrm{c} + \Delta c(N)$.

It has been found that in quantum dots with large electron number Δc is typically small and therefore

$$V_H(N) = (N-1)V_c$$

is a good approximation. The proportionality to the electron number is consistent with the fact that the Hartree potential contains the electron density in the dot which has a normalized spatial distribution that does not change appreciably if a single electron is added to more than 100 electrons already present (Koopman's theorem).

(3) It turns out that, in limited intervals of gate voltages, the lever arms of relevant states are independent of the index n of the states. As a consequence, a constant gate lever arm α_i can be defined which is independent of n.

Considering only the contribution to the Hartree energy which is proportional to N, neglecting the exchange interaction and assuming state-independent lever arms results in the so-called *constant interaction model*.

Taking these approximations in the Hartree–Fock model of the quantum dot, we obtain the electrochemical potential in the constant interactions model

$$\mu_N = \epsilon_N^{(0)} - e\sum_i \alpha_i(\phi_i - \phi_i^{(0)}) + (N-1)V_c, \tag{18.26}$$

where the quantities V_c (charging energy) and α_i (lever arms of the gates) are constants. The single-particle energies ϵ_n may be derived from a model potential of noninteracting electrons.

The constant interactions are based on a number of severe approximations. Nevertheless it has proven to form a useful basis for capturing the underlying physics of many of the observed effects.

18.2.10 Configuration interaction, exact diagonalization

The configuration interaction method, also called exact diagonalization, requires a lot of computer power for calculating quantum states and can therefore only be applied for relatively small electron numbers (up to around 10). We choose a basis of single-particle states, such as the Fock–Darwin states of the two-dimensional harmonic oscillator in the perpendicular magnetic field. For the calculation of the N-particle eigenstates we form N-particle Slater determinants $|\,\phi_i\rangle$ (the so-called configurations) which obey the symmetry requirement for fermionic systems upon particle exchange. If the calculation is based on a finite number of $n > N$ single-particle wave functions, we obtain

$$\begin{pmatrix} n \\ N \end{pmatrix}$$

N-particle wave functions. The hamiltonian can then be written and diagonalized in matrix form with elements $H_{ij} = \langle \phi_i | H | \phi_j \rangle$. The resulting eigenstates are linear combinations of Slater determinants, i.e, they have the form

$$\psi = \sum_i c_i \, | \, \phi_i \rangle \ .$$

The advantage of this approach compared to the Hartree or Hartree–Fock methods is that correlation effects between electrons can be correctly described. These correlations are more important the more Slater determinants contribute to a particular state, i.e., the broader the distribution of the $|c_i|^2$ plotted as a function of the index i. The disadvantage of the method is the large numerical effort. The accuracy of the calculations can be checked by inspecting changes of the calculated spectra as the number of single-particle basis states is changed.

18.3 Electronic transport through quantum dots

So far we have discussed the states of completely isolated quantum dots without considering the weak tunneling coupling to source and drain contacts quantitatively. We ascribed the current flow qualitatively to an alignment between electrochemical potentials in the source, the dot, and the drain contact. In the following we will consider the tunneling transport through quantum dots quantitatively.

18.3.1 Resonant tunneling

Electron transport through quantum dot structures is closely related to the resonant tunneling phenomenon. In order to discuss those properties of a resonant tunneling structure that are relevant for quantum dots, we consider one-dimensional model systems in which noninteracting particles are scattered at a double barrier structure. They offer the advantage that they expose the involved physics without complications due to higher dimensionality.

Feynman-paths and resonant tunneling. The resonant tunneling problem in one dimension can, in full analogy to a Fabry–Perot interferometer in optics, be described as the interference of partial waves (see Fig. 18.28). In a quantum dot the electron waves play the role of the partial waves of light in the optical interferometer, and the role of the semitransparent mirrors is played by the tunneling barriers connecting

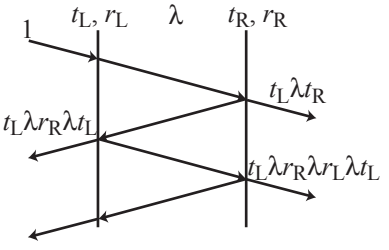


Fig. 18.28 Schematic of interfering paths reflected back and forth between two semitransparent mirrors, which are in quantum dot physics replaced by the tunneling barriers coupling the dot to source and drain.

the dot to source and drain contacts. In this picture the total transmission amplitude is

$$\begin{aligned} t &= t_L\lambda t_R + t_L\lambda r_R\lambda r_L\lambda t_R + t_L\lambda r_R\lambda r_L\lambda r_R\lambda r_L\lambda t_R + \cdots \\ &= t_L\lambda t_R \sum_{n=0}^{\infty} (r_R\lambda^2 r_L)^n \\ &= \frac{t_L\lambda t_R}{1 - r_L\lambda^2 r_R}. \end{aligned}$$

This result corresponds to eq. (18.30) that will later be obtained from the double delta barrier problem.

For further discussion of this result we wish to transform it to a different form. The expression in the denominator of the transmission can be written as $1 - |r_L||r_r|\exp(i\theta)$, where $\theta = \arg(r_\mathrm{L}) + \arg(r_\mathrm{R}) + 2\arg(\lambda)$ is the phase that an electron accumulates on a round trip between the barriers. Also the numerator of the expression for the transmission contains a factor $\exp(i\theta/2)$, and we write it as $|t_\mathrm{L}||t_\mathrm{R}|\exp(i\theta/2 + i\alpha)$, where $\alpha = \arg(t_\mathrm{L}) + \arg(t_\mathrm{R}) - \arg(r_\mathrm{L})/2 - \arg(r_\mathrm{R})/2$. Then the transmission amplitude is given as

$$\begin{aligned} t = \frac{|t_L||t_R|e^{i(\theta/2+\alpha)}}{1 - |r_L||r_R|e^{i\theta}} &= \frac{|t_L||t_R|e^{i\alpha}}{e^{-i\theta/2} - |r_L||r_R|e^{i\theta/2}} \\ &= \frac{|t_L||t_R|}{1 - |r_L||r_R|} \frac{e^{i\alpha}}{\cos(\theta/2) - i\frac{1+|r_L||r_R|}{1-|r_L||r_R|}\sin(\theta/2)}. \end{aligned}$$

Introducing the weakly energy-dependent coupling strength

$$\gamma = \frac{1 - |r_L||r_R|}{1 + |r_L||r_R|} < 1$$

we obtain for the transmission

$$t = \frac{|t_L||t_R|e^{i\alpha}}{1 - |r_L||r_R|} \frac{1}{\cos(\theta/2) - i\gamma^{-1}\sin(\theta/2)}. \tag{18.27}$$

As the considered energy is increased, the angle θ increases and the denominator $D = \cos(\theta/2) - i\gamma^{-1}\sin(\theta/2)$ describes an elliptic curve in the complex plane, extending between $[-1, +1]$ along the real axis and $[-\gamma^{-1}, +\gamma^{-1}]$ along the imaginary axis as plotted in Fig. 18.29. The magnitude of the transmission (18.27) is maximum when the magnitude of D is minimum, i.e., resonances occur for $\theta = 2\pi p$, where p is an integer. At resonance, the magnitude of the second factor in (18.27) is always one because there, $(\sin\theta/2) = 0$, and $\cos(\theta/2) = \pm 1$. Therefore the prefactor is the amplitude of the resonances.

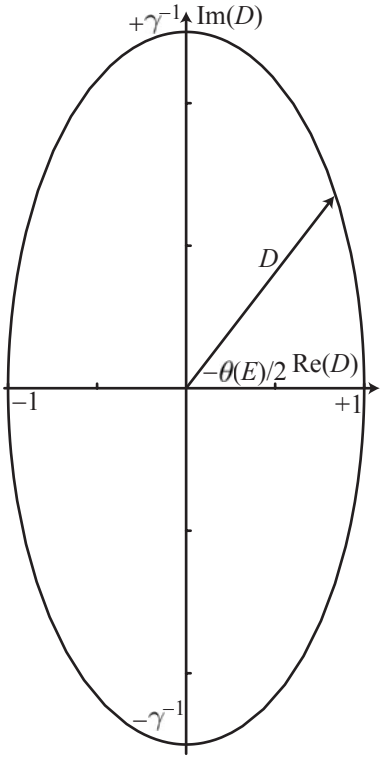


Fig. 18.29 Value of the denominator D in the expression for the resonant transmission amplitude following an ellipse with increasing angle θ.

Lorentz approximation. We now consider a very important implication of this result called the *Lorentz approximation*. It is an approximation for the transmission $t(E)$ near a resonance, given that $|t_\mathrm{L}|, |t_\mathrm{R}|$ are much smaller than one implying that r_L and $\mathrm{r_R}$ are close to one.

To this end we realize that close to a resonance at energy E_p, we can expand

$$\theta(E) \approx 2\pi p + \left.\frac{d\theta(E)}{dE}\right|_{E=E_p} (E - E_p).$$

Since the second term is small by definition, we have close to a resonance $\cos[\theta(E)/2] \approx \pm 1$ (plus sign for even p, minus for odd), and

$$\sin[\theta(E)/2] \approx \sin\left[\pi p + \frac{1}{2}\left.\frac{d\theta(E)}{dE}\right|_{E=E_p} (E - E_p)\right] \approx \pm\frac{1}{2}\left.\frac{d\theta(E)}{dE}\right|_{E=E_p} (E - E_p),$$

where the plus sign refers to even p, and the minus sign to odd p. With the resonance line width Γ_p defined by

$$\frac{1}{\Gamma_p} = \frac{1}{4\gamma}\left.\frac{d\theta(E)}{dE}\right|_{E=E_p}$$

and introducing this approximation into the expression for the transmission amplitude (18.27) we obtain

$$t = \pm\frac{|t_L||t_R|e^{i\alpha}}{1 - |r_L||r_R|}\frac{1}{1 - i(E - E_p)/(\Gamma_p/2)} := \frac{t_0}{1 - i(E - E_p)/(\Gamma_p/2)}. \quad (18.28)$$

This expression is the transmission amplitude in the Lorentz approximation. Comparing the denominator of this expression with the denominator D depicted in Fig. 18.29, we realize that the approximation is better the more extended the ellipse is along the imaginary axis, i.e., the smaller γ. This quantity, however, becomes smaller the larger the reflection at the two barriers, i.e., the smaller is their transmission. This makes clear that the Lorentz approximation is valid for $|t_\mathrm{L}|, |t_\mathrm{R}| \ll 1$, and therefore $\mathcal{T}_\mathrm{L}, \mathcal{T}_\mathrm{R} \ll 1$.

For symmetric barriers, i.e., $|r_\mathrm{L}| = |r_\mathrm{R}|$, and $|t_\mathrm{L}| = |t_\mathrm{R}|$, we have $t_0 = \pm 1$, and therefore

$$t_\mathrm{symm} = \pm\frac{1}{1 - i(E - E_p)/(\Gamma_p/2)},$$

because $|t_{\mathrm{L/R}}|^2 = 1 - |r_{\mathrm{L/R}}|^2$. The magnitude of the transmission is for symmetric barriers on the resonance ($E = E_p$) equal to one.

The transmission probability for asymmetric barriers is a lorentzian

$$\mathcal{T} = \frac{|t_0|^2}{1 + (E - E_p)^2/(\Gamma_p/2)^2},$$

as depicted in Fig. 18.30.

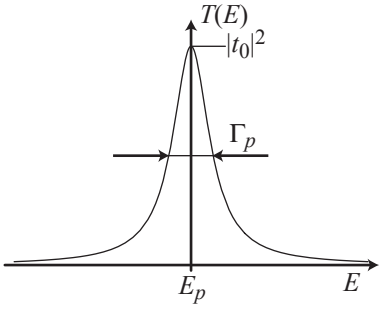


Fig. 18.30 Transmission probability for a lorentzian resonance. The resonance is centered around the energy E_p, has a width of 2Γ and an amplitude of $|t_0|^2$.

If we have a closer look at the amplitude $|t_0|^2$ of the resonance and again use the fact that $\mathcal{T}_{\mathrm{L/R}} = |t_{\mathrm{L/R}}|^2 = 1 - |r_{\mathrm{L/R}}|^2$ we can write

$$|t_0|^2 = \frac{\mathcal{T}_{\mathrm{L}}\mathcal{T}_{\mathrm{R}}}{(1 - \sqrt{(1-\mathcal{T}_{\mathrm{L}})(1-\mathcal{T}_{\mathrm{R}})})^2}.$$

If we make—in the spirit of the Lorentz approximation—the assumption that $\mathcal{T}_{\mathrm{L}}, \mathcal{T}_{\mathrm{R}} \ll 1$, we can expand the denominator and find

$$|t_0|^2 = \frac{4\mathcal{T}_{\mathrm{L}}\mathcal{T}_{\mathrm{R}}}{(\mathcal{T}_{\mathrm{L}} + \mathcal{T}_{\mathrm{R}})^2}.$$

For a given average transmission $\mathcal{T} \equiv (\mathcal{T}_{\mathrm{L}} + \mathcal{T}_{\mathrm{R}})/2$, this amplitude has a maximum as a function of $\mathcal{T}_{\mathrm{L}} - \mathcal{T}_{\mathrm{R}}$ at the symmetry point $\mathcal{T}_{\mathrm{L}} = \mathcal{T}_{\mathrm{R}}$. This result tells us that conductance resonances show their maximum amplitude if the tunneling coupling to the source and drain barriers is symmetric.

Further insight about the resonance line width can be obtained by inspecting γ for the case $\mathcal{T}_{\mathrm{L}}, \mathcal{T}_{\mathrm{R}} \ll 1$. We find

$$\gamma = \frac{1 - |r_L||r_R|}{1 + |r_L||r_R|} \approx \frac{\mathcal{T}_{\mathrm{R}} + \mathcal{T}_{\mathrm{L}}}{4}.$$

As a consequence, the quantity Γ_p contains additive contributions of the two barriers because

$$\Gamma_p = 4\gamma \left(\left.\frac{d\theta(E)}{dE}\right|_{E=E_p} \right)^{-1} \approx (\mathcal{T}_{\mathrm{R}} + \mathcal{T}_{\mathrm{L}}) \left(\left.\frac{d\theta(E)}{dE}\right|_{E=E_p} \right)^{-1} = \Gamma_p^{(\mathrm{L})} + \Gamma_p^{(\mathrm{R})},$$

where

$$\Gamma_p^{(\mathrm{L/R})} := \mathcal{T}_{\mathrm{L/R}}(E) \left(\left.\frac{\partial\theta(E)}{\partial E}\right|_{E=E_p} \right)^{-1}.$$

With these results we can finally write the transmission for a sequence of resonances in the Lorentz approximation as

$$\mathcal{T} = \sum_p \frac{\Gamma_p^{(\mathrm{L})}\Gamma_p^{(\mathrm{R})}}{\Gamma_p^{(\mathrm{L})} + \Gamma_p^{(\mathrm{R})}} \frac{\Gamma_p}{(\Gamma_p/2)^2 + (E - E_p)^2} := \sum_p \frac{\Gamma_{\mathrm{L}}(p)\Gamma_{\mathrm{R}}(p)}{\Gamma_{\mathrm{L}}(p) + \Gamma_{\mathrm{R}}(p)} \mathcal{L}_p[E - E_p], \quad (18.29)$$

where $\Gamma_p = \Gamma_p^{(\mathrm{L})} + \Gamma_p^{(\mathrm{R})}$. In this approximation, the energy dependence of the $\Gamma_{\mathrm{L/R}}$ is usually neglected.

According to the approximations that we made for arriving at the Lorentz approximation, the resulting expression for the transmission is only valid near the resonances in the case of weak tunneling coupling. The total tunneling coupling is now described by the two parameters $\Gamma_p^{(\mathrm{L/R})}$ which are the rates with which a particle would tunnel

out of the potential well. We can see this if we consider the definition of $\theta = 2\arg(\lambda) + \arg(r_\mathrm{L}) + \arg(r_\mathrm{R})$ and assume a typical behavior $\arg(\lambda) = kL$, where k is the wave vector between the barriers and L is the barrier separation. The derivative $\partial\theta/\partial E$ contains the contribution $2L(\partial E/\partial k)^{-1}$. The partial derivative of the energy with respect to the wave vector is (up to a factor $\hbar$) equal to the group velocity of the electron. The distance $2L$ between the barriers divided by this velocity is the mean time interval between collisions of the electron with one of the barriers. This means that the electron attempts to tunnel out of the potential well with a rate

$$\nu = \frac{1}{2L}\frac{1}{\hbar}\frac{\partial E}{\partial k}.$$

However, each attempt is only successful with the probability $\mathcal{T}_{\mathrm{L/R}}$ such that $\Gamma_{\mathrm{L/R}}/\hbar = \nu\mathcal{T}_{\mathrm{L/R}}$ is the tunneling rate through the left (right) barrier. In this discussion we have neglected the (typically weak) energy dependence of the reflection coefficients r_L and r_R.

More complicated models for resonant tunneling with islands of higher dimensionality or more realistic tunneling barriers show conceptually the same behavior. However, such models are usually more tedious to solve because the Schrödinger equation has to be solved in three dimensions for a given quantum dot potential. The tunneling coupling can often only be considered approximately in a perturbative treatment. Often the transfer hamiltonian approach is used for this purpose.

Coherent tunneling through two delta scatterers. After having seen the concept of resonant tunneling in analogy to a Fabry–Perot interferometer, we consider resonant tunneling within an analytically solvable problem consisting of two δ-potentials. The scattering potential is given by $U(x) = \sum_{i=1}^{2} U_i\delta(x - x_i)$. The two barriers enclose a region of length $L = |x_1 - x_2|$ in which particles can move freely. The model is schematically depicted in Fig. 18.31.

Transmission of the double barrier structure. When we look for the transmission of the double barrier structure we can build on the

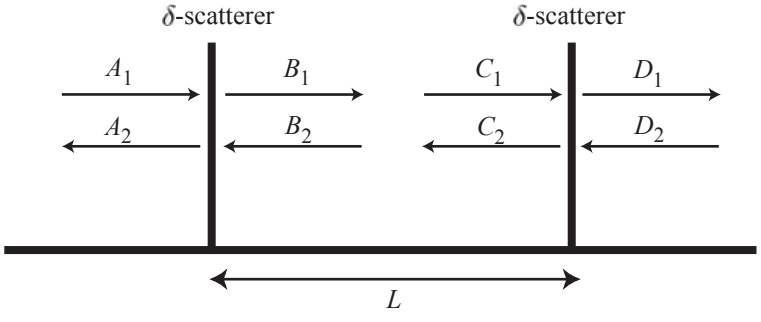


Fig. 18.31 One-dimensional scattering potential consisting of two δ-potentials having a separation L. The quantities $A_1, A_2, B_1, B_2, \ldots$ are the amplitudes of the right- and left-moving plane waves that are solutions of the scattering potential.

single-barrier result discussed in section 12.1. We know the wave functions in the three regions to the left of the first barrier, between the two barriers, and to the right of the second barrier:

$$\begin{aligned}\psi_L(x) &= a_1 e^{ikx} + a_2 e^{-ikx} \\ \psi_Z(x) &= b_1 e^{ikx} + b_2 e^{-ikx} \\ \psi_R(x) &= c_2 e^{ikx} + c_2 e^{-ikx}\end{aligned}$$

We define the amplitudes of the wave functions to the left and right of the individual barriers as $A_1 = a_1 e^{ikx_1}$, $A_2 = a_2 e^{-ikx_2}$, $B_1 = b_1 e^{ikx_1}$, $B_2 = b_2 e^{-ikx_1}$, $C_1 = b_1 e^{ikx_2}$, $C_2 = b_2 e^{-ikx_2}$, $D_1 = c_1 e^{ikx_2}$, and $D_2 = c_2 e^{-ikx_2}$. Then we can write the boundary conditions at the two scatterers as

$$\begin{aligned}A_1 + A_2 &= B_1 + B_2 \\ ik(B_1 - B_2) - ik(A_1 - A_2) &= \gamma(A_1 + A_2) \\ C_1 + C_2 &= D_1 + D_2 \\ ik(D_1 - D_2) - ik(C_1 - C_2) &= \gamma'(C_1 + C_2).\end{aligned}$$

The first pair of equations is identical with the boundary conditions of the single barrier (see section 12.1). The second pair has the same structure as the first and we can write

$$\begin{aligned}\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} &= T_k \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \\ \begin{pmatrix} D_1 \\ D_2 \end{pmatrix} &= T'_k \begin{pmatrix} C_1 \\ C_2 \end{pmatrix},\end{aligned}$$

where the matrix elements of T'_k, the transfer matrix of the second barrier, are determined by γ'. Propagation of waves between the two scatterers is described by the equation

$$\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \underbrace{\begin{pmatrix} e^{ikL} & 0 \\ 0 & e^{-ikL} \end{pmatrix}}_{P_k} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix},$$

with the propagator P_k. In order to simplify the notation we define $\lambda := e^{ikL}$.

Using the matrices T_k, T'_k, and P_k we can write the transmission problem as

$$\mathbf{D} = T_k P_k T'_k \mathbf{A} := M_k \mathbf{A},$$

where M_k is the total transfer matrix of the system. In order to calculate the matrix M_k we first determine the product

$$T_k P_k = \begin{pmatrix} \alpha & \beta^\star \\ \beta & \alpha^\star \end{pmatrix} \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^\star \end{pmatrix} = \begin{pmatrix} \alpha\lambda & \beta^\star\lambda^\star \\ \beta\lambda & \alpha^\star\lambda^\star \end{pmatrix},$$

and find

$$\begin{aligned}M_k = T_k P_k T'_k &= \begin{pmatrix} \alpha\lambda & \beta^\star\lambda^\star \\ \beta\lambda & \alpha^\star\lambda^\star \end{pmatrix} \begin{pmatrix} \alpha' & \beta'^\star \\ \beta' & \alpha'^\star \end{pmatrix} \\ &= \begin{pmatrix} \alpha\lambda\alpha' + \beta^\star\lambda^\star\beta' & \alpha\lambda\beta'^\star + \beta^\star\lambda^\star\alpha'^\star \\ \beta\lambda\alpha' + \alpha^\star\lambda^\star\beta' & \beta\lambda\beta'^\star + \alpha^\star\lambda^\star\alpha'^\star \end{pmatrix}.\end{aligned}$$

This matrix has the same structure as the transfer matrix for the single barrier in eq. (12.1), i.e.,

$$\begin{pmatrix} D_1 \\ D_2 \end{pmatrix} = \begin{pmatrix} c_1 e^{ikx_2} \\ c_2 e^{-ikx_2} \end{pmatrix} = \begin{pmatrix} \mu & \nu^\star \\ \nu & \mu^\star \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} \mu & \nu^\star \\ \nu & \mu^\star \end{pmatrix} \begin{pmatrix} a_1 e^{ikx_1} \\ a_2 e^{-ikx_1} \end{pmatrix}.$$

In order to calculate the transmission amplitude we let $a_1 = 1$, $a_2 = r$, $c_1 = t$, and $c_2 = 0$ and obtain

$$\begin{pmatrix} te^{ikx_2} \\ 0 \end{pmatrix} = \begin{pmatrix} \mu & \nu^\star \\ \nu & \mu^\star \end{pmatrix} \begin{pmatrix} e^{ikx_1} \\ re^{-ikx_1} \end{pmatrix}.$$

It follows that

$$\begin{aligned} r &= -\frac{\nu}{\mu^\star} e^{2ikx_1} \\ t &= \frac{|\mu|^2 - |\nu|^2}{\mu^\star} e^{-ikL}. \end{aligned}$$

In order to find our final result, we express μ and ν by the transmission coefficients of the single barriers. This leads to

$$\begin{aligned} \mu &= \frac{1 + f r_L^\star r_R^\star (\lambda^\star)^2}{t_L^\star \lambda^\star t_R^\star} = \frac{1 - r_L^\star r_R^\star (\lambda^\star)^2}{t_L^\star \lambda^\star t_R^\star} \\ \nu &= -\frac{r_L \left(1 + f^{-1} \frac{r_L r_R}{|r_L|^2} \lambda^2\right)}{t_L \lambda t_R} = -\frac{r_L \left(1 - \frac{r_L r_R}{|r_L|^2} \lambda^2\right)}{t_L \lambda t_R}, \end{aligned}$$

where we have used the relation

$$f = \frac{t_L^\star r_L}{t_L r_L^\star} = -\frac{(1 + i\gamma/2k)}{(1 - i\gamma/2k)} \frac{i\gamma/2k/(1 + i\gamma/2k)}{i\gamma/2k/(1 - i\gamma/2k)} = -1.$$

With these relations it can be shown that $|\mu|^2 - |\nu|^2 = 1$. This is a general result of time-reversal symmetry. The transmission coefficient of the whole structure is therefore $e^{-ikL}/\mu^\star$ leading to

$$t = \frac{t_L t_R}{1 - r_L r_R \lambda^2} = \frac{t_L t_R}{1 - |r_L||r_R| e^{i\theta}}, \tag{18.30}$$

where $\theta = 2kL + \arg(r_L) + \arg(r_R)$ is the phase that the partial wave picks up on a round trip between the barriers with two reflections. The transmission probability of the double barrier structure is therefore

$$\mathcal{T} = \frac{\mathcal{T}_L \mathcal{T}_R}{1 + \mathcal{R}_L \mathcal{R}_R - 2\sqrt{\mathcal{R}_L \mathcal{R}_R} \cos\theta}. \tag{18.31}$$

For the reflection coefficient we find

$$r = \frac{1}{r_L^\star} \frac{|r_L|^2 - |r_L||r_R| e^{i\theta}}{1 - |r_L||r_R| e^{i\theta}} e^{2ikx_1},$$

and the reflection probability is

$$\mathcal{R} = \frac{\mathcal{R}_L + \mathcal{R}_R - 2\sqrt{\mathcal{R}_L \mathcal{R}_R}\cos\theta}{1 + \mathcal{R}_L \mathcal{R}_R - 2\sqrt{\mathcal{R}_L \mathcal{R}_R}\cos\theta}.$$

It can be verified also that in this case, as for the single barrier, $\mathcal{T} + \mathcal{R} = 1$.

The total transmission has a maximum at those energies where $\theta = 2\pi n$ (n integer), because the transmission and reflection probabilities of the single barriers depend only weakly on energy.

Figure 18.32(a) shows the transmission for identical barriers with $\gamma L = \gamma' L = 10$. Transmission resonances indicate the occurrence of resonant tunneling. The resonances arise at particular values of kL, i.e., at particular energies. The value of the transmission is one at these resonances. With increasing energy the resonance widths increase and, as a result, the background transmission between the resonances increases steadily with energy. This is related to the fact that the transmission of the individual barriers increases with increasing energy (see Fig. 12.2). As a consequence, an electron placed at time $t = 0$ between the barriers will tunnel within a time span Δt into one of the two leads. The finite probability of finding an electron between the barriers leads according to the time–energy uncertainty relation $\Delta E \Delta t > h$ to a level broadening $\Delta E \sim h/\Delta t$ between the barriers and therefore to the broadening of the transmission resonances.

Density of states between the barriers. Within the same model we can calculate the wave functions $\psi_k(x)$ in the region between the two δ-scatterers. The quantity

$$\rho_k = 2\int_0^L dx\, |\psi_k(x)|^2 / (2\pi)$$

is the density of states on this 'island'. The factor of 2 in front of the integral accounts for the fact that scattering states impinging from the left and from the right contribute to the density of states. Figure 18.32(b) shows the density of states for $\gamma L = 1$. It exhibits maxima at the same values of kL, i.e., at the same energies as the transmission. With increasing energy the peaks in the density of states broaden and their amplitude is reduced, in contrast to the amplitude of the transmission resonances. These peaks in the density of states resemble the discrete energy spectrum of a system with opaque barriers (quantum well). Indeed, the peaks in the density of states become sharper if the strength of the δ-scatterers increases, or if the energy of the incident particle decreases. The width of the density of states peaks depend again on the lifetime of a particle between the barriers, according to Heisenberg's uncertainty relation.

Number of particles between the barriers. If we fill the quantum states of this one-dimensional noninteracting system with fermions up

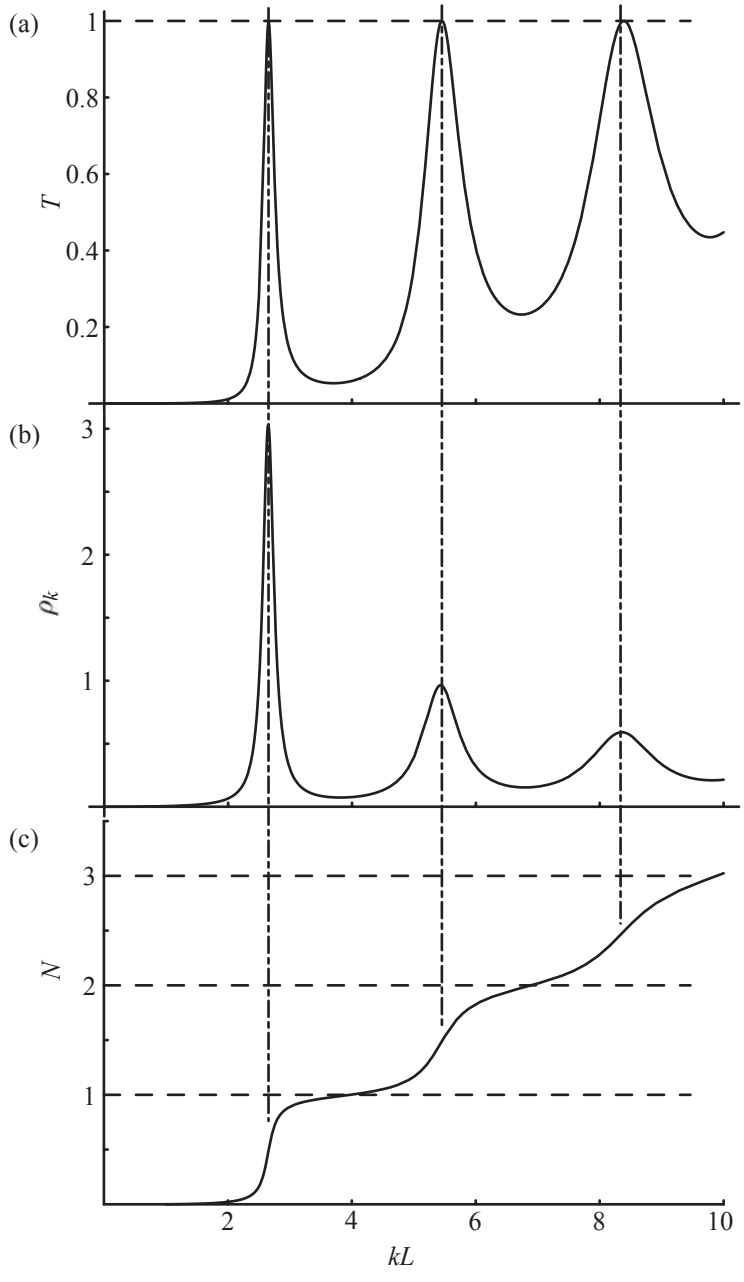


Fig. 18.32 Transmission T (a), density of states ρ_k (b), and number of particles N (c) calculated for the double δ-barrier problem with parameter $\gamma = 10$. In (a), the dashed horizontal line indicates $T = 1$; in (c) horizontal dashed lines represent integer electron numbers N. The vertical dash–dotted lines show that resonances in the transmission (a) and the density of states (b) occur at the same energy (kL), and these give rise to the step-like increase of the particle number N on the island.

to a given Fermi wave vector k_{F}, we obtain a the number of particles N on the island from

$$N = \int_0^{k_{\mathrm{F}}} \rho_k dk.$$

This number of particles N is shown in Fig. 18.32(c) as a function of $k_{\mathrm{F}}L$. Steps can be seen at integer numbers of N leading to a monotonic and continuous increase with energy. The steps are more pronounced at small energies (small kL) and tend to wash out at larger energies. It turns out that the steps are very pronounced as long as the transmission of the individual barriers is much smaller than one (cf., Fig. 12.2). In the opposite case, when the transmission of the individual barriers is close to one ($1 - T \ll 1$), the steps disappear completely. The step-like character of this function is the more pronounced the larger γ, i.e., the weaker the tunneling coupling of the island to the leads. This one-dimensional model of an island shows that the number of particles on the island is quantized if it is weakly tunneling coupled to the leads. This quantization is crucial for the Coulomb blockade effect, although the latter involves strong interactions between electrons on the island.

This result can also be expressed using the conductance G of a single barrier. The number of particles on an island is quantized only if the tunneling coupling through the individual barriers leads to a single-barrier conductance $G \ll e^2/h$. Otherwise, the quantization of the particle number on the island breaks down. This result is in agreement with the experimental findings shown in Fig. 18.7.

Resonant tunneling current at small bias voltage

Using the Landauer–Büttiker theory, the conductance G of a resonant tunneling structure can be determined for small source–drain voltage U_{SD} and finite temperature. In the Lorentz approximation (18.29) we obtain

$$G = \frac{e^2}{h} \sum_p \frac{\Gamma_p^{(\mathrm{L})} \Gamma_p^{(\mathrm{R})}}{\Gamma_p^{(\mathrm{L})} + \Gamma_p^{(R)}} \int dE \mathcal{L}_p(E - E_p) \left(\frac{\partial f}{\partial E} \right). \tag{18.32}$$

Here we assume that the sum over the resonant levels p includes the two spin orientations. Again we would like to emphasize here that the presented theory does not contain any interaction effects between electrons on the island. It can therefore not describe the Coulomb blockade phenomenon. The expression given above is nevertheless similar to measurements in the Coulomb blockade. The current exhibits sharp lorentzian resonances if the temperature $k_{\mathrm{B}}T$ is small compared to the width Γ_p. If $k_{\mathrm{B}}T \gg \Gamma_p$, then the resonance is thermally broadened. In this situation the conductance is given by

$$G = \frac{e^2}{h} \frac{1}{4k_{\mathrm{B}}T} \sum_p \frac{\Gamma_p^{(\mathrm{L})} \Gamma_p^{(\mathrm{R})}}{\Gamma_p^{(\mathrm{L})} + \Gamma_p^{(R)}} \frac{1}{\cosh^2[(E_p - E_{\mathrm{F}})/2k_{\mathrm{B}}T]}. \tag{18.33}$$

The expression can be regarded as the sum of individual transmission channels, each contributing an amount

$$G_p = \frac{e^2}{h}\frac{1}{4k_{\mathrm{B}}T}\frac{\Gamma_p^{(\mathrm{L})}\Gamma_p^{(\mathrm{R})}}{\Gamma_p^{(\mathrm{L})}+\Gamma_p^{(R)}}\frac{1}{\cosh^2[(E_p-E_{\mathrm{F}})/2k_{\mathrm{B}}T]} \tag{18.34}$$

to the total conductance.

18.3.2 Sequential tunneling

For resonant tunneling the coherence of contributing paths in the Feynman description was essential. We will now turn our attention to a model for tunneling transport through quantum dots in which only the existence and occupation of discrete states is important. We will set up a rate equation which describes the occupation statistics of the states. In this rate equation approach, the interaction between electrons in the dot can be introduced seamlessly. The approach is based on a perturbation treatment of the tunneling coupling. Its validity is therefore limited to the situation where the tunneling coupling is by far the smallest energy scale in the problem. Using the rate equation approach for the description of electron transport through quantum dots has been introduced in Averin *et al.*, 1991, and in Beenakker, 1991. Particular examples for its use and important special cases can, for example, be found in Bonet *et al.*, 2002.

Theoretical treatment

System hamiltonian and notation. We describe the system of source and drain lead, and quantum dot by the hamiltonian

$$H = H_0 + H^{(1)} = H_{\mathrm{S}} + H_{\mathrm{D}} + H_{\mathrm{d}} + H^{(1)},$$

where H_{S} and H_{D} describe noninteracting electrons in the source and drain lead, respectively, H_{d} describes the electron motion in the (interacting) quantum dot, and $H^{(1)}$ is the tunneling coupling between the leads and the dot. The operator H_0 is the sum of the lead and dot hamiltonians which we assume to be diagonalized.

In the following, single-particle states in the leads are labeled with the Greek letters $\Lambda\lambda$ where $\Lambda = \mathrm{S}$ (source) or $\Lambda = \mathrm{D}$ (drain) states in the dot with the Greek letter δ. Many-body states of the contacts are labeled $\ell, \ell', \ldots$ together with S/D, those in the dot $d, d', \ldots$. States of the whole system are labeled, for example, $n = (\mathrm{S}\ell, \mathrm{D}\ell', d)$.

Perturbation expansion in the tunneling coupling. We describe states of the system with the density matrix ρ, because we are dealing with a many-body system with the leads acting as thermodynamic baths. The dynamics is then governed by the von Neumann equation $i\hbar\partial_t\rho(t) = [H, \rho(t)]$ which we transform into the interaction picture in which each operator A is transformed according to

$$A(t) = e^{iH^{(0)}t/\hbar}Ae^{-iH^{(0)}t/\hbar}.$$

The resulting von Neumann equation in the interaction picture is

$$i\hbar\partial_t\rho(t) = [H^{(1)}(t), \rho(t)]. \tag{18.35}$$

This equation is naturally suited for a perturbation expansion of $\rho(t)$. For this purpose we decompose the density matrix into the sum

$$\rho(t) = \sum_i \rho^{(i)}(t) \tag{18.36}$$

with corrections of order i. If we insert this expansion into eq. (18.35) and treat $H^{(1)}(t)$ as a first-order term, we can collect terms of increasing order on both sides of the equation and find for the i^{th} order the recursive relation

$$i\hbar\partial_t\rho^{(i)}(t) = \begin{cases} 0 & \text{for } i = 0 \\ \left[H^{(1)}(t), \rho^{(i-1)}(t)\right] & \text{for } i > 0 \end{cases}. \tag{18.37}$$

The lowest-order density matrix is therefore stationary.

The tunneling hamiltonian. We now specify the tunneling hamiltonian in second quantization

$$H^{(1)} = H^{(1)}_{\text{S}} + H^{(1)}_{\text{D}} = \sum_{\substack{\delta\lambda \\ \Lambda=\text{S,D}}} \left(t^{(\Lambda)}_{\delta\lambda} d^\dagger_\delta a_{\Lambda\lambda} + t^{(\Lambda)\star}_{\delta\lambda} d_\delta a^\dagger_{\Lambda\lambda} \right), \tag{18.38}$$

where $d^\dagger_\delta$ (d_δ) create (annihilate) a particle in state δ in the dot, and $a^\dagger_{\Lambda\lambda}$ ($a_{\Lambda\lambda}$) create (annihilate) a particle in state λ in lead Λ. The tunneling coupling to each lead can be decomposed into a tunneling-in and a tunneling-out process, i.e.,

$$H^{(1)}_\Lambda = H^{(\text{in})}_\Lambda + H^{(\text{out})}_\Lambda \tag{18.39}$$

with

$$H^{(\text{in})}_\Lambda = \sum_{\delta\lambda} t^{(\Lambda)}_{\delta\lambda} d^\dagger_\delta a_{\Lambda\lambda} \quad \text{and} \quad H^{(\text{out})}_\Lambda = \sum_{\delta\lambda} t^{(\Lambda)\star}_{\delta\lambda} d_\delta a^\dagger_{\Lambda\lambda}.$$

We note here that $H^{(\text{in})}_\Lambda = H^{(\text{out})\dagger}_\Lambda$.

An expression for the current. The operator of the current is calculated from the operator of the change of charge in the source lead in time, i.e., from

$$I(t) = -|e|\partial_t N_{\text{S}}(t),$$

where the number operator $N_{\text{S}}(t) = \sum_\lambda a^\dagger_{\text{S}\lambda}(t) a_{\text{S}\lambda}(t)$. Its time evolution is governed by Heisenberg's equation

$$\partial_t N_{\text{S}}(t) = \frac{i}{\hbar}[H(t), N_{\text{S}}(t)] = \frac{i}{\hbar}[H^{(1)}_{\text{S}}(t), N_{\text{S}}(t)]. \tag{18.40}$$

Here, we have used the fact that the number operator of the source lead commutes with the whole hamiltonian, except $H^{(1)}_{\text{S}}$. We can work out

the commutator required in eq. (18.40) by looking at the matrix elements of the involved operators. The matrix elements of the number operator are given by $N_{\mathrm{S},nm}(t) = \delta_{nm} n_{\mathrm{S},n}$, where $n_{\mathrm{S},n}$ is the number of electrons in the source lead in the system state n. The required commutator has matrix elements

$$\left[H_{\mathrm{S}}^{(1)}(t), N_{\mathrm{S}}(t)\right]_{nm} = e^{i\omega_{nm}t} H_{\mathrm{S},nm}^{(1)} \left(n_{\mathrm{S},m} - n_{\mathrm{S},n}\right) = e^{i\omega_{nm}t} \left(H_{\mathrm{S},nm}^{(\mathrm{in})} - H_{\mathrm{S},nm}^{(\mathrm{out})}\right),$$

and we therefore obtain for the desired commutator the operator expression $\left[H_{\mathrm{S}}^{(1)}(t), N_{\mathrm{S}}(t)\right] = H_{\mathrm{S}}^{(\mathrm{in})}(t) - H_{\mathrm{S}}^{(\mathrm{out})}(t)$.

The operator for the current is then given by the intuitive expression

$$I(t) = -|e| \frac{i}{\hbar} \left(H_{\mathrm{S}}^{(\mathrm{in})}(t) - H_{\mathrm{S}}^{(\mathrm{out})}(t)\right).$$

The expectation value for the current is

$$\langle I(t) \rangle = -|e| \langle \partial_t N_{\mathrm{S}}(t) \rangle = -|e| \frac{i}{\hbar} \mathrm{trace} \left\{ \left[H_{\mathrm{S}}^{(\mathrm{in})}(t) - H_{\mathrm{S}}^{(\mathrm{out})}(t)\right] \rho(t) \right\}.$$

The perturbation expansion of the density matrix (18.36) will give a corresponding perturbation expansion of the expectation value of the current

$$\langle I(t) \rangle = \sum_{i=0}^{\infty} \left\langle I^{(i)}(t) \right\rangle$$

with $\left\langle I^{(0)}(t) \right\rangle = 0$ (absence of equilibrium currents) and

$$\left\langle I^{(i)}(t) \right\rangle = -|e| \frac{i}{\hbar} \mathrm{trace} \left\{ \left[H_{\mathrm{S}}^{(\mathrm{in})}(t) - H_{\mathrm{S}}^{(\mathrm{out})}(t)\right] \rho^{(i)}(t) \right\}. \qquad (18.41)$$

The current is therefore the sum of corrections of higher order in the time evolution of the density matrix.

Density matrix in first order. With eqs (18.37) and (18.41) we have the basis for the calculation of the current in first order in the tunneling coupling. For the first-order correction of the equilibrium density matrix we find according to eq. (18.37)

$$i\hbar \partial_t \rho^{(1)}(t) = [H^{(1)}(t), \rho^{(0)}(t)] = e^{iH^{(0)}t/\hbar} [H^{(1)}, \rho^{(0)}] e^{-iH^{(0)}t/\hbar}. \quad (18.42)$$

For making further progress we need to specify $\rho^{(0)}$. We assume an equilibrium thermal density matrix

$$\rho^{(0)} = \rho^{(\mathrm{S})} \otimes \rho^{(\mathrm{D})} \otimes \rho^{(\mathrm{d})}$$

with $\rho_{\ell,\ell'}^{(\Lambda)} = \delta_{\ell,\ell'} p_\ell^{(\Lambda)}$, and $\rho_{d,d'}^{(\mathrm{d})} = \delta_{d,d'} p_d$. The $p_\ell^{(\Lambda)}$ (p_d) are thermal equilibrium probabilities for states $(\Lambda\ell)$ and (d) in the leads and dot, respectively.

If m, n, and k denote states of the entire system under the influence of $H^{(0)}$, the commutator in eq. (18.42) has matrix elements

$$[H^{(1)},\rho^{(0)}]_{mn} = \sum_k \left(H^{(1)}_{mk}\rho^{(0)}_{kn} - \rho^{(0)}_{mk}H^{(1)}_{kn} \right) = H^{(1)}_{mn}\left(p_n - p_m\right),$$

where we have abbreviated $p_n = p^{(\mathrm{S})}_{\ell} p^{(\mathrm{D})}_{\ell'} p_d$, and $p_m = p^{(\mathrm{S})}_{\ell''} p^{(\mathrm{D})}_{\ell'''} p_{d'}$. With this result, eq. (18.42) becomes

$$i\hbar\partial_t \rho^{(1)}_{mn}(t) = e^{i\omega_{mn}t} H^{(1)}_{mn}(p_n - p_m)$$

with $\omega_{mn} = E_{mn}/\hbar = (E_m - E_n)/\hbar$. Time integration leads to

$$\rho^{(1)}_{mn}(t) = -g(E_{mn},t) H^{(1)}_{mn}(p_n - p_m). \tag{18.43}$$

At this point we have defined the function

$$g(E_{nm},t) := \frac{e^{i\omega_{nm}t}-1}{E_{nm}} = 2ie^{i\omega_{nm}t/2}\frac{\sin(E_{nm}t/2\hbar)}{E_{nm}}. \tag{18.44}$$

The function has a finite value $it/\hbar$ at $E_{nm} = 0$ and oscillates as a function of energy. The oscillation amplitude decays with energy as $1/E_{nm}$ and it has the symmetry property that $g(E_{nm},t) = -g^\star(E_{mn},t)$.

Tunneling current in first order. In order to evaluate the current in first order according to eq. (18.41) we write it out in matrix notation, insert the density matrix in first order from eq. (18.43), and obtain

$$\left\langle I^{(1)}(t)\right\rangle = -|e|\frac{i}{\hbar}\left\{ \sum_{mn} g(E_{nm},t)[H^{(\mathrm{in})}_{\mathrm{S},nm} - H^{(\mathrm{out})}_{\mathrm{S},nm}]H^{(1)}_{mn}(p_n - p_m)\right\}. \tag{18.45}$$

The product of the expression in square brackets with $H^{(1)}_{mn}$ can be further simplified considering the properties of the tunneling hamiltonian (18.38):

$$\begin{aligned}[H^{(\mathrm{in})}_{\mathrm{S},nm} - H^{(\mathrm{out})}_{\mathrm{S},nm}]H^{(1)}_{mn} &= H^{(\mathrm{in})}_{\mathrm{S},nm}H^{(\mathrm{out})}_{\mathrm{S},mn} - H^{(\mathrm{out})}_{\mathrm{S},nm}H^{(\mathrm{in})}_{\mathrm{S},mn} \\ &= \left|H^{(\mathrm{in})}_{\mathrm{S},nm}\right|^2 - \left|H^{(\mathrm{out})}_{\mathrm{S},nm}\right|^2 \end{aligned} \tag{18.46}$$

The first matrix element squared represents a tunneling-in rate, the second a tunneling-out rate. Only one of the two can be nonzero for a given pair of states (nm).

Realizing that the diagonal terms in the sum over (mn) are zero, we can simplify the expression for the current by transforming eq. (18.45) as $\sum_{mn} f_{nm} = \sum_{mn}(f_{nm} + f_{mn})/2$ into

$$\begin{aligned}\left\langle I^{(1)}(t)\right\rangle &= \frac{|e|}{\hbar}\left\{ \sum_{mn}\frac{\sin\omega_{nm}t}{E_{nm}}\left[\left|H^{(\mathrm{in})}_{\mathrm{S},nm}\right|^2 - \left|H^{(\mathrm{out})}_{\mathrm{S},nm}\right|^2\right](p_n - p_m)\right\} \\ &= \frac{2|e|}{\hbar}\sum_{mn}\frac{\sin(E_{nm}t/\hbar)}{E_{nm}}\left[\left|H^{(\mathrm{out})}_{\mathrm{S},mn}\right|^2 - \left|H^{(\mathrm{in})}_{\mathrm{S},mn}\right|^2\right]p_n \end{aligned} \tag{18.47}$$

The interpretation of this expression for the current is as follows: Starting from state n, which is found with probability p_n, the system evolves into state m under the influence of the tunneling hamiltonian which transfers an electron from the source lead into the dot or from the dot into the source lead. The sign of the contribution to the tunneling current depends on the direction of tunneling. The time-dependent prefactor decays rapidly as the energy difference E_{nm} increases and in the limit of large times it becomes a delta function. In this limit of large times t we therefore have

$$\left\langle I^{(1)} \right\rangle = |e| \frac{2\pi}{\hbar} \sum_{mn} \delta(E_{nm}) \left[\left| H^{(\mathrm{out})}_{\mathrm{S},mn} \right|^2 - \left| H^{(\mathrm{in})}_{\mathrm{S},mn} \right|^2 \right] p_n, \quad (18.48)$$

expressing the fact that tunneling conserves the total energy of the system. The tunneling-out term in this expression is given by

$$\sum_{mn} \delta(E_{nm}) \left| H^{(\mathrm{out})}_{\mathrm{S},mn} \right|^2 p_n = \sum_{dd'} p_d \sum_{\delta\lambda} \left| t^{(\mathrm{S})}_{\delta\lambda} \right|^2 |\langle d'| \, d_\delta \, |d\rangle|^2 \, \delta(\mu_{dd'} - \epsilon_\lambda) \sum_{\ell\ell'} \left| \langle \ell'| \, a^\dagger_{\mathrm{S},\lambda} \, |\ell\rangle \right|^2 p^{(\mathrm{S})}_\ell$$

and the tunneling-in term is

$$\sum_{mn} \delta(E_{nm}) \left| H^{(\mathrm{in})}_{\mathrm{S},mn} \right|^2 p_n = \sum_{dd'} p_d \sum_{\delta\lambda} \left| t^{(\mathrm{S})}_{\delta\lambda} \right|^2 \left| \langle d'| \, d^\dagger_\delta \, |d\rangle \right|^2 \delta(\mu_{d'd} - \epsilon_\lambda) \sum_{\ell\ell'} |\langle \ell'| \, a_{\mathrm{S},\lambda} \, |\ell\rangle|^2 p^{(\mathrm{S})}_\ell .$$

The last matrix elements on the right-hand side can be further simplified, because the leads are noninteracting. We find

$$\sum_{\ell\ell'} |\langle \ell| \, a_{\mathrm{S}\lambda} \, |\ell'\rangle|^2 p^{(\mathrm{S})}_{\ell'} = f_{\mathrm{S}}(\epsilon_\lambda), \quad \sum_{\ell\ell'} \left| \langle \ell| \, a^\dagger_{\mathrm{S}\lambda} \, |\ell'\rangle \right|^2 p^{(\mathrm{S})}_{\ell'} = 1 - f_{\mathrm{S}}(\epsilon_\lambda),$$
$$\sum_{\ell\ell'} |\langle \ell| \, a_{\mathrm{S}\lambda} \, |\ell'\rangle|^2 p^{(\mathrm{S})}_{\ell} = 1 - f_{\mathrm{S}}(\epsilon_\lambda), \quad \sum_{\ell\ell'} \left| \langle \ell| \, a^\dagger_{\mathrm{S}\lambda} \, |\ell'\rangle \right|^2 p^{(\mathrm{S})}_{\ell} = f_{\mathrm{S}}(\epsilon_\lambda),$$

where $f_{\mathrm{S}}(\epsilon_\lambda)$ is the Fermi–Dirac equilibrium distribution function in the source lead, evaluated at the energy ϵ_λ of state λ. Inserting all the above results into eq. (18.48) gives

$$\left\langle I^{(1)} \right\rangle = -|e| \sum_{dd'} p_d \left[\Gamma^{(\mathrm{S})}_{dd'} f_{\mathrm{S}}(\mu_{d'd}) - \Gamma^{(\mathrm{S})}_{d'd} (1 - f_{\mathrm{S}}(\mu_{dd'})) \right] \quad (18.49)$$

with the tunneling rates for lead Λ defined as

$$\Gamma^{(\Lambda)}_{dd'} = \frac{2\pi}{\hbar} \sum_{\delta\lambda} |\langle d| \, d_\delta \, |d'\rangle|^2 \left| t^{(\Lambda)}_{\delta\lambda} \right|^2 \delta(|\mu_{dd'}| - \epsilon_\lambda). \quad (18.50)$$

The rate $\Gamma^{(\Lambda)}_{dd'}$ is nonzero only for pairs of dot states (d, d') for which the electron number $N_d = N_{d'} - 1$. Equation (18.49) together with the rates

(18.50) are the central result for the current. The expressions for the rates look very similar to the standard Fermi's golden rule, but there is an additional matrix element measuring the overlap between state d', say, an N-electron state with an electron added in state δ, and the $N+1$-electron state d. Suppose, for example, that state d and (d' + electron in state δ) have orthogonal spin components. In such a case, tunneling-in would be completely suppressed, although energy might be conserved and the tunneling matrix element be nonzero. Tunneling-in and -out therefore occurs with particular rates that ensure energy conservation and spin conservation. Equation (18.49) tells us that the total current is the result of a competition between tunneling-in from the source into the dot and tunneling-out from the dot into the source contact. Every dot state d can contribute to a tunneling current provided it is occupied (factor p_d), and provided the target state in the lead is empty (tunneling-out) or occupied (tunneling-in).

Rate equations for the occupation statistics. The above calculation of the tunneling current through a quantum dot leaves us with the open question of how to determine the dot occupation factors p_d. We will now show that these can be obtained as the stationary solution of a rate equation for the reduced density matrix of the quantum dot. Treating the system in 0^{th} order does not give such a rate equation, as is evident from eq. (18.37). It turns out that tracing out the lead states in the first-order result in eq. (18.43) does not give the rate equation either. In order to derive the rate equation we have to calculate the second-order correction to the density matrix of the whole system, and then trace out the states of the leads. The second order correction to the density matrix is found from eq. (18.37) to be

$$i\hbar\partial_t\rho^{(2)}(t) = [H^{(1)}(t), \rho^{(1)}(t)].$$

Going again to matrix notation in eigenstates of the unperturbed hamiltonian $H^{(0)}$ this equation becomes

$$i\hbar\partial_t\rho^{(2)}_{mn}(t) = \sum_k \left[e^{i\omega_{mk}t} H^{(1)}_{mk}\rho^{(1)}_{kn}(t) - \rho^{(1)}_{mk}(t) H^{(1)}_{kn} e^{i\omega_{kn}t} \right].$$

Inserting the result eq. (18.43) for the first-order correction we obtain

$$i\hbar\partial_t\rho^{(2)}_{mn}(t) = -\sum_k H^{(1)}_{mk}H^{(1)}_{kn} \left[\left(e^{i\omega_{mn}t} - e^{i\omega_{mk}t}\right) \frac{p_n - p_k}{E_{kn}} - \left(e^{i\omega_{mn}t} - e^{i\omega_{kn}t}\right) \frac{p_k - p_m}{E_{mk}} \right],$$

and after time integration

$$\rho^{(2)}_{mn}(t) = \sum_k H^{(1)}_{mk}H^{(1)}_{kn} \left\{ [g(E_{mn},t) - g(E_{mk},t)] \frac{p_n - p_k}{E_{kn}} - [g(E_{mn},t) - g(E_{kn},t)] \frac{p_k - p_m}{E_{mk}} \right\}. \quad (18.51)$$

We use the notation $|m\rangle = |\ell'\tau' d'\rangle$ and $|n\rangle = |\ell\tau d\rangle$, where ℓ and ℓ' denote states in the source, whereas τ and τ' denote states in the drain contact. We can trace out the states ℓ, ℓ', and τ, τ' of the contacts in eq. (18.51) and obtain the reduced density matrix of the dot

$$\rho^{(2)}_{\mathrm{d},d'd}(t) = \sum_{\ell\tau k} H^{(1)}_{\ell\tau d',k} H^{(1)}_{k,\ell\tau d} \left[\left(g(\mu_{d'd},t) - g(E_{\ell\tau d',k},t)\right) \frac{p_{\ell\tau d} - p_k}{E_{k,\ell\tau d}} \right.$$
$$\left. - \left(g(\mu_{d'd},t) - g(E_{k,\ell\tau d},t)\right) \frac{p_k - p_{\ell\tau d'}}{E_{\ell\tau d',k}} \right]$$

with the time derivative

$$\partial_t \rho^{(2)}_{\mathrm{d},d'd}(t) = \frac{i}{\hbar} \sum_{\ell\tau k} H^{(1)}_{\ell\tau d',k} H^{(1)}_{k,\ell\tau d} \left[\left(e^{i\omega_{d'd}t} - e^{i\omega_{\ell\tau d',k}t}\right) \frac{p_{\ell\tau d} - p_k}{E_{k,\ell\tau d}} \right.$$
$$\left. - \left(e^{i\omega_{d'd}t} - e^{i\omega_{k,\ell\tau d}t}\right) \frac{p_k - p_{\ell\tau d'}}{E_{\ell\tau d',k}} \right].$$

The diagonal elements of the reduced density matrix obey the equations

$$\partial_t \rho^{(2)}_{\mathrm{d},dd}(t) = -\frac{2}{\hbar} \sum_{\ell\tau k} \left| H^{(1)}_{\ell\tau d,k} \right|^2 (p_{\ell\tau d} - p_k) \frac{\sin(E_{k,\ell\tau d}t/\hbar)}{E_{k,\ell\tau d}}.$$

On long time scales $t/\hbar \to \infty$ this equation goes into Fermi's golden rule result

$$\partial_t \rho^{(2)}_{\mathrm{d},dd}(t) = \frac{2\pi}{\hbar} \sum_{\substack{\ell\tau \\ \ell'\tau' d'}} \left| H^{(1)}_{\ell\tau d,\ell'\tau' d'} \right|^2 \delta(E_{\ell'\tau' d',\ell\tau d})(p_{\ell'\tau' d'} - p_{\ell\tau d}).$$

This equation is equivalent to the rate equation

$$\partial_t p_d(t) = \sum_{d'} \left[W(d,d') - \sum_{d''} W(d'',d)\delta_{dd'} \right] p_{d'}, \tag{18.52}$$

where $p_d(t) \equiv \rho^{(2)}_{\mathrm{d},dd}(t)$, and the

$$W(d,d') = \frac{2\pi}{\hbar} \sum_{\substack{\ell\tau \\ \ell'\tau'}} \left| H^{(1)}_{\ell\tau d,\ell'\tau' d'} \right|^2 \delta(E_{\ell'\tau',\ell\tau} - \mu_{dd'}) p_{\ell'} p_{\tau'}$$

are the transition rates from state d' to d, or d to d', respectively. It just remains to evaluate these transition rates further using eq. (18.38) in the same fashion as in the case of the tunneling current in lowest order. The square of the tunneling matrix element is

$$\begin{aligned} \left| H^{(1)}_{n,m} \right|^2 &= \left| H^{\mathrm{S}}_{nm} \right|^2 + \left| H^{\mathrm{D}}_{nm} \right|^2 \\ &= \left| H^{\mathrm{S,in}}_{nm} \right|^2 + \left| H^{\mathrm{S,out}}_{nm} \right|^2 + \left| H^{\mathrm{D,in}}_{nm} \right|^2 + \left| H^{\mathrm{D,out}}_{nm} \right|^2 \end{aligned}$$

and after some more algebra we arrive at

$$\begin{aligned} W(d,d') &= \Gamma^{(\mathrm{S})}_{d'd} f_{\mathrm{S}}(\mu_{dd'}) + \Gamma^{(\mathrm{S})}_{dd'} \left[1 - f_{\mathrm{S}}(\mu_{d'd})\right] \\ &\quad + \Gamma^{(\mathrm{D})}_{d'd} f_{\mathrm{D}}(\mu_{dd'}) + \Gamma^{(\mathrm{D})}_{dd'} \left[1 - f_{\mathrm{D}}(\mu_{d'd})\right], \end{aligned} \tag{18.53}$$

where we have used the definition of the tunneling rates from eq. (18.50).

Equations (18.52) and (18.53) are the main results that allow the determination of the population of quantum dot states. Their interpretation is as follows. If the quantum dot under consideration is in a state d' with N electrons, it can decay into a state d with $N-1$ electrons, if an electron tunnels from the dot into the source, or the drain reservoir [second and fourth term in eq. (18.53)]. It may alternatively decay into a state d with $N+1$ electrons, if an electron tunnels from one of the reservoirs into the dot [first and third term in eq.(18.53)]. For each of these transitions there exists a transition rate $W(d,d')$ from the initial state d' into the final state d.

We can find a *stationary* probability p_d which describes the probability that the quantum dot is found in a particular state d. The probabilities p_d are determined by the rate equation (18.52), where we set $\partial_t p_d(t) = 0$ in order to obtain the stationary solutions, i.e.,

$$0 = \sum_{d'} \left[W(d,d') - \sum_{d''} W(d'',d)\delta_{dd'} \right] p_{d'}. \tag{18.54}$$

The expression in square brackets forms a matrix with determinant zero. Therefore the additional requirement that the probability distribution p_d be normalized, i.e.,

$$\sum_d p_d = 1, \tag{18.55}$$

is needed to produce a unique solution together with the above rate equations (18.54).

Summary. The above eqs (18.54) and (18.55) lead to a stationary distribution function describing the occupation of quantum dot states which is not the well-known Fermi–Dirac distribution for a gas of free noninteracting fermions. The required transition rates are defined in eq. (18.53) which can be evaluated with the help of the definition (18.50). Having the dot occupation distribution, the current can be calculated from eq. (18.49).

Example: System with two quantum dot states

As the simplest example, we consider a quantum dot in which only two ground states differing by one in electron number are relevant for the electronic transport. This situation is given if the temperature is much smaller than the lowest excitations of the quantum dot above the two involved ground states, and, of course, smaller than the charging energy. We denote the energies of the two ground states with $E_\mathrm{d}(1)$ and $E_\mathrm{d}(2)$, with $E_\mathrm{d}(2) > E_\mathrm{d}(1)$. Further, let $N_1 = N-1$ and $N_2 = N$. From the rate equation (18.54) and the normalization condition (18.55) we find

the occupation probabilities of the two states

$$\begin{aligned} p_1 &= \frac{W(1,2)}{W(2,1)+W(1,2)} \\ p_2 &= \frac{W(2,1)}{W(2,1)+W(1,2)}. \end{aligned}$$

Inserting the expressions (18.53) for the rates, and realizing that the Γ_{12}^{Δ} are the only nonzero rates, leads to

$$\begin{aligned} p_1 &= \frac{\Gamma_{12}^{(\mathrm{S})}[1-f_{\mathrm{S}}(\mu_N)]+\Gamma_{12}^{(\mathrm{D})}[1-f_{\mathrm{D}}(\mu_N)]}{\Gamma_{12}^{(\mathrm{S})}+\Gamma_{12}^{\mathrm{D}}} \\ p_2 &= \frac{\Gamma_{12}^{(\mathrm{S})}f_{\mathrm{S}}(\mu_N)+\Gamma_{12}^{(\mathrm{D})}f_{\mathrm{D}}(\mu_N)}{\Gamma_{12}^{(\mathrm{S})}+\Gamma_{12}^{\mathrm{D}}}, \end{aligned}$$

where the Fermi functions have to be evaluated at the energy $\mu_N = E(2) - E(1)$. A special case arises if no bias voltage is applied between source and drain. In this case $f_{\mathrm{S}} = f_{\mathrm{D}} \equiv f$, and we obtain

$$\begin{aligned} p_1 &= 1-f \\ p_2 &= f. \end{aligned}$$

This result tells us that the occupation of the state with one excess electron on the dot goes to zero sharply, as soon as μ_N is shifted above the electrochemical potentials (Fermi energies) in the contacts, e.g., by the application of a plunger gate voltage. At the same time, the occupation probability of the state with zero excess electrons on the dot rises.

The current given by eq. (18.49) is the difference of two contributions, the current created by an electron tunneling from the dot into the source, and the current created by the opposite process, i.e.,

$$I = +|e|\left\{p_2\Gamma_{12}^{(\mathrm{S})}[1-f_{\mathrm{S}}(\mu_N)] - p_1\Gamma_{12}^{(\mathrm{S})}f_{\mathrm{S}}(\mu_N)\right\}.$$

Inserting the expressions for the occupation probabilities we obtain

$$I = -\frac{|e|}{h}\frac{\Gamma_{12}^{(\mathrm{S})}\Gamma_{12}^{(\mathrm{D})}}{\Gamma_{12}^{(\mathrm{S})}+\Gamma_{12}^{(\mathrm{D})}}\left[f_{\mathrm{S}}(\mu_N)-f_{\mathrm{D}}(\mu_N)\right]. \tag{18.56}$$

At zero source–drain voltage we have $\mu_{\mathrm{S}} = \mu_{\mathrm{D}}$ and therefore also $f_{\mathrm{S}}(\mu_N) = f_{\mathrm{D}}(\mu_N)$, and the current is zero. Figure 18.33 shows two $I(V_{\mathrm{SD}})$ curves calculated according to eq. (18.56). It was assumed that $\mu_{\mathrm{S}} = eV_{\mathrm{SD}}/2$, $\mu_{\mathrm{D}} = -eV_{\mathrm{SD}}/2$, and $\mu_N = |e|(\alpha_{\mathrm{S}}-\alpha_{\mathrm{D}})V_{\mathrm{SD}}/2 - |e|\alpha_{\mathrm{G}}V_{\mathrm{G}} + \mathrm{const}$. A characteristic property of the curve in the Coulomb blockade (off resonance) is the exponential suppression of the current for small source–drain voltages. In contrast, on a conductance resonance (on resonance), the trace is linear for small source–drain voltages. Comparison with the measured curve in Fig. 18.2 shows that the characteristic behavior of the measurement is quite well represented by our simple model. In the following we will derive the special case of linear transport from eq. (18.56).

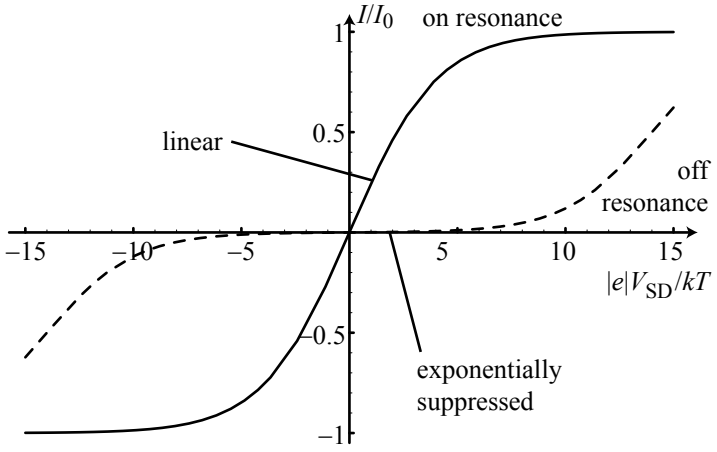


Fig. 18.33 Current–voltage characteristics of a quantum dot in the Coulomb blockade regime (off resonance) and on a conductance peak (on resonance), calculated according to eq. (18.56).

For small source–drain voltages (linear response) we can, for example, Taylor-expand f_S up to first order:

$$f_S(\mu_N) = f_D(\mu_N) + \left.\frac{df_D(\mu_N)}{d\mu_D}\right|_{\mu_S=\mu_D} (\mu_S - \mu_D)$$
$$= f_D(\mu_N) - \frac{|e|V_{SD}}{4k_BT\cosh^2[(\mu_N - \mu_D)/2k_BT]}.$$

The current is then given by

$$I = \frac{e^2}{h}\frac{\Gamma_{12}^{(S)}\Gamma_{12}^{(D)}}{\Gamma_{12}^{(S)} + \Gamma_{12}^{(D)}}\frac{1}{4k_BT\cosh^2[(\mu_N - \mu_D)/2k_BT]}V_{SD}$$

and the conductance is

$$G = G_0\cosh^{-2}[(\mu_N - \mu_D)/2k_BT]$$

with

$$G_0 = \frac{e^2}{h}\frac{\Gamma_{12}^{(S)}\Gamma_{12}^{(D)}}{\Gamma_{12}^{(S)} + \Gamma_{12}^{(D)}}\frac{1}{4k_BT}.$$

This result gives us insights into several experimental findings. A peak in the linear conductance will always be found as a function of gate voltages whenever the electrochemical potential μ_N in the quantum dot $\mu_N = \mu_D$ (cf., Fig. 18.9). This resonance condition can be controlled through the gate voltages as described by eq. (18.25). We can therefore write the function describing a conductance resonance as

$$G = \frac{G_0}{\cosh^2[\alpha_{pg}(V_{pg} - V_{res})/2k_BT]}. \qquad (18.57)$$

The peak is thermally broadened, and decays exponentially with increasing distance from the maximum, as shown in Fig.18.34. From the width

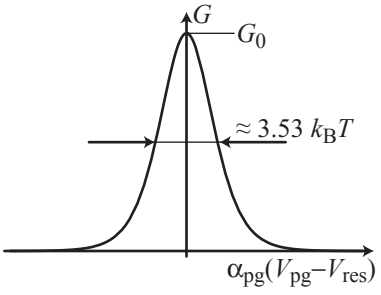


Fig. 18.34 Shape of a conductance resonance according to eq. (18.57) in the single-level tunneling regime, if the tunneling coupling is much smaller than temperature.

of such a resonance the electron temperature can be experimentally determined. Within the above model, the height of conductance peaks decays in inverse proportion to the temperature. We can write for the prefactor G_0 giving the height of the resonance

$$G_0 = \left(\frac{h}{e^2} \frac{4k_\mathrm{B}T}{\Gamma_{12}^{(\mathrm{S})}} + \frac{h}{e^2} \frac{4k_\mathrm{B}T}{\Gamma_{12}^{(\mathrm{D})}} \right)^{-1} := (R_\mathrm{S} + R_\mathrm{D})^{-1} ,$$

i.e., it can be interpreted as adding the two tunneling resistances R_S and R_D in series.

Remarkable about the result for the conductance in eq. (18.57) is the fact that it corresponds exactly to the expression for the resonant tunneling current for one transmission channel in eq. (18.34). It therefore turns out to be an interesting question as to whether the current through a quantum dot is resonant and coherent, or sequential and incoherent. This question cannot be answered with measurements of the Coulomb blockade effect alone. Interference measurements have to be designed, where quantum dots are embedded in ring geometries. Such measurements have shown that a part of the current through a quantum dot can indeed be coherent.

If the simple two-state model discussed above is applicable, we talk about the *single-level transport regime*. If we take a value of 30 μeV for the typical single-particle level spacing, it requires experiments at temperatures well below 300 mK to reach the single-level transport regime.

Figure 18.35(a) shows measured conductance peaks of a quantum dot fabricated from a GaAs/AlGaAs heterostructure. At small plunger gate voltage ($V_\mathrm{pg} \approx 282.5$ mV), eq. (18.57) gives an excellent fit [Fig. 18.35(b)]. At larger plunger gate voltages, a convolution of a lorentzian resonance with the derivative of the Fermi–Dirac distribution gives a very good fit [Fig. 18.35(c)].

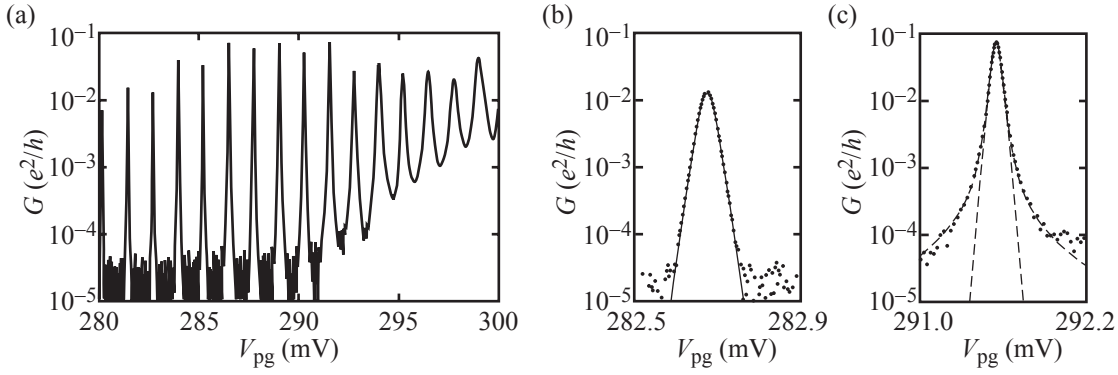


Fig. 18.35 (a) Coulomb blockade measurements made on a quantum dot in a GaAs/AlGaAs heterostructure. (b) Fit of a particular conductance resonance at small plunger gate voltage with the thermally broadened function in eq. (18.57). (c) At larger plunger gate voltages where the source and drain tunneling coupling is stronger, a fit with a thermally broadened lorentzian is better than eq. (18.57). (Reprinted with permission from Foxman *et al.*, 1993. Copyright 1993 by the American Physical Society.)

Beyond the single level transport regime

Experimentally, situations can arise where the reduction of the problem to two dot states, i.e., one electrochemical potential μ_N, is not possible. This is the case, for example, if excitations of the quantum dot are very close in energy to the ground states. As a result, the shapes of conductance resonances in linear response can become more complicated and, beyond the energy difference μ_N between ground states, other energy differences $\mu_N^{(n,m)}$ occur [cf., eq.(18.9)].

In systems beyond the single-level transport regime, the shape of conductance resonances can be calculated on the basis of the rate equation (18.54) and the equation for the current (18.49). Analytic results can be obtained for certain limiting cases if the constant interaction model for the quantum dot is used (Beenakker, 1991). The main assumption is that the energy level broadening Γ caused by the tunneling coupling to source and drain contacts is much smaller than the thermal energy $k_\mathrm{B}T$, i.e., $\Gamma \ll k_\mathrm{B}T$. It turns out that the conductance is the sum over contributions of individual transport channels with a given electron number N in the dot, where transport takes place through the single-particle level p:

$$G = \sum_N \sum_p G_{N,p}.$$

The contribution of the energy level ϵ_p to the conductance of the dot with N electrons is given by

$$G_{N,p} = \frac{e^2}{k_\mathrm{B}T} \frac{\Gamma_p^{(\mathrm{S})}\Gamma_p^{(\mathrm{D})}}{\Gamma_p^{(\mathrm{S})} + \Gamma_p^{(\mathrm{D})}} F_{eq}(\epsilon_p, N)\,[1 - f(\mu_N)]\,,$$

where the $\Gamma_p^{(\mathrm{S})}$ and $\Gamma_p^{(\mathrm{D})}$ describe the tunneling coupling of the energy level ϵ_p to source (S) and drain (D). The function $f(E)$ is the Fermi–Dirac distribution in the source (drain) contact. The function $F_\mathrm{eq}(\epsilon_p, N)$ represents the statistical probability that the energy level ϵ_p of the N-electron dot is thermally occupied.

At sufficiently low temperatures, $k_\mathrm{B}T \ll V_\mathrm{c}$, only one particular electron number N contributes to electron transport. Furthermore if $k_\mathrm{B}T$ is smaller than the separation of neighboring energy levels ϵ_p, only one single-particle energy level ϵ_p contributes. This is the case of the single-level transport regime introduced above. For this very important special case the above formula for the conductance simplifies to eq. (18.57).

18.3.3 Higher order tunneling processes: cotunneling

Figure 18.36(b) shows dI/dV_SD diamond measurements obtained from a GaAs quantum dot fabricated by AFM lithography. A number of excited quantum dot states can be seen outside the diamonds. The sum of charging energy and single-particle level spacing is about 2 meV, indicating that the dot is quite small. Typical single-particle level spacings are around 200 µeV. Within the diamonds, structure can be seen

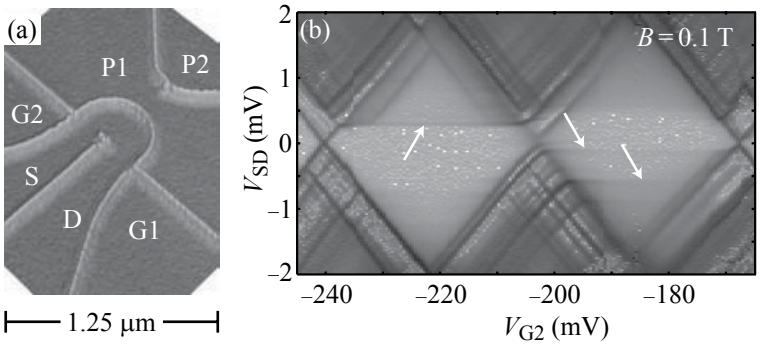


Fig. 18.36 (a) AFM image of the quantum dot sample used for the measurement of the differential conductance in (b). (b) Measurements of $dI/dV_{\rm SD}$ in the plane of plunger gate and source–drain voltage. A number of excited states can be seen outside the Coulomb blockade diamonds. Inside the diamonds, the signatures of inelastic cotunneling and elastic cotunneling through excited states can be seen (white arrows). (Reprinted with permission from Schleser *et al.*, 2005. Copyright 2005 by the American Physical Society.)

in $dI/dV_{\rm SD}$ which cannot be explained within the models of electron transport through quantum dots discussed so far. The corresponding so-called cotunneling processes will be discussed in the following.

Figure 18.37 shows a schematic Coulomb blockade diamond with energy schemes for different transport processes. Processes 1 to 4 can be described within the sequential elastic tunneling picture discussed earlier. Beyond that we distinguish elastic cotunneling processes (process 5) and inelastic cotunneling processes (process 6).

In an elastic cotunneling process, an electron tunnels from source to drain via a virtual intermediate nonresonant state in the dot. Although this virtual state is higher in energy than the electron's energy in the source and drain contact, the total process is energy-conserving, and the dot is found in its ground state before and after the electron transfer. In an inelastic cotunneling process, two electrons tunnel in a correlated fashion and the electron is left in an excited state (or it starts from an excited state and ends up in the ground state). All these processes occur at finite source–drain voltages.

Elastic cotunneling. Elastic cotunneling (processes 5 in Fig. 18.37) dominates the conductance between conductance resonances if the tunneling coupling is sufficiently strong such that first-order tunneling is no longer a good approximation. As an example of the physics involved, we consider the quantum mechanical transition amplitude for tunneling from source to drain for the case of tunneling via the virtual states μ_N and μ_{N+1}, as depicted in Fig. 18.38.

In lowest order they are given by the sum of the two alternative processes (a) and (b)

$$t_{\tau\ell} = \frac{t^{\star}_{\ell d} t_{\tau d}}{\epsilon_\tau - \mu_N} + \frac{t^{\star}_{\ell d'} t_{\tau d'}}{\mu_{N+1} - \epsilon_\ell},$$

where ℓ numbers a state in the source contact, τ a state in the drain, and d and d' are the two intermediate dot states relating to μ_N and μ_{N+1}, respectively. According to Fermi's golden rule and the condition that

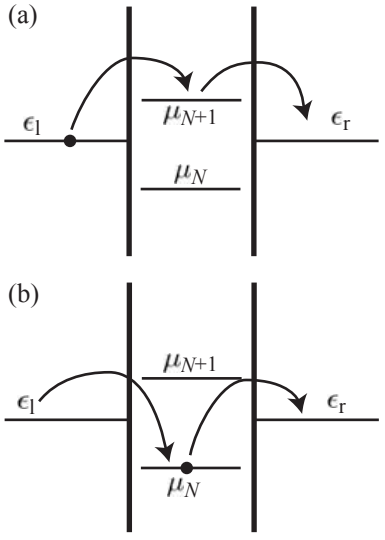


Fig. 18.38 Two tunneling processes contributing to elastic cotunneling.

the initial state of the electron in the source contact has to be occupied whereas the final state in the drain has to be empty, the total tunneling rate is given by

$$W_{\mathrm{D}\leftarrow\mathrm{S}} = \frac{2\pi}{\hbar}\sum_{\ell\tau}|t_{\tau\ell}|^2\,\delta(\epsilon_\tau-\epsilon_\ell)f_{\mathrm{S}}(\epsilon_\ell)[1-f_{\mathrm{D}}(\epsilon_\tau)].$$

The summation goes over all states in the source and drain contact. The tunneling rate is a sum of three terms, one of which is an interference term, because the tunneling amplitude is the sum of two terms. Assuming that the tunneling rates are independent of energy over a small source–drain voltage interval, we obtain at zero temperature

$$W_{\mathrm{DS}} = \frac{2\pi}{\hbar}\left[\frac{\Gamma_{\mathrm{S}}(N)\Gamma_{\mathrm{D}}(N)}{(\mu_{\mathrm{D}}-\mu_N)(\mu_{\mathrm{S}}-\mu_N)}+\frac{\Gamma_{\mathrm{S}}(N+1)\Gamma_{\mathrm{D}}(N+1)}{(\mu_{\mathrm{D}}-\mu_{N+1})(\mu_{\mathrm{S}}-\mu_{N+1})}\right]eV_{\mathrm{SD}} + \frac{\ln\left(\frac{\mu_{N+1}-\mu_{\mathrm{D}}}{\mu_{\mathrm{D}}-\mu_N}\right)-\ln\left(\frac{\mu_{N+1}-\mu_{\mathrm{S}}}{\mu_{\mathrm{S}}-\mu_N}\right)}{(\mu_{N+1}-\mu_N)(\mu_{\mathrm{S}}-\mu_{\mathrm{D}})}.$$

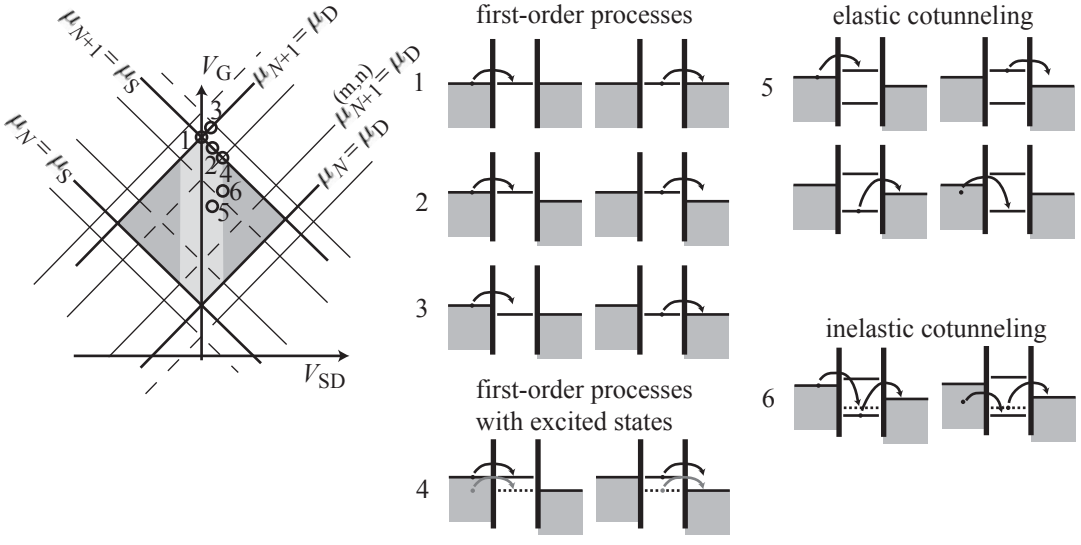


Fig. 18.37 Schematic Coulomb blockade diamond together with the energy diagram of selected tunneling processes of first and second order (cotunneling). (1) Sequential tunneling with source–drain voltage close to zero (linear response). (2) Plunger gate voltage reduced and source–drain voltage increased such that $\mu_{N+1}=\mu_{\mathrm{S}}$. This condition is fulfilled along the line labeled correspondingly. (3) Starting from (1) the plunger gate voltage and the source–drain voltage were increased such that $\mu_{N+1}=\mu_{\mathrm{D}}$. This condition is fulfilled along the line labeled correspondingly. (4) We went from (1) along the line $\mu_{N+1}=\mu_{\mathrm{S}}$ to larger source–drain voltage until an excited state transition $\mu_N^{(m,n)}$ became just possible, such that two transport channels are in the bias window. (5) Two elastic cotunneling processes using virtual intermediate states of the quantum dot. The second process can be regarded as the hole analogue to the first. Such processes are possible everywhere within the diamond. (6) Inelastic cotunneling process where the quantum dot is left in an excited state and the tunneling electron loses the same amount of energy (left-hand diagram). The required energy is therefore supplied by the applied voltage, i.e., $\mu_N^{(m,n)}-\mu_N=eV_{\mathrm{SD}}$. This condition defines the borderline between light and dark gray regions in the diamond. At the diamond edge this line ends, where an excited state line continues outside the diamond. Such inelastic processes are possible in the dark gray regions of the diamond. The right-hand diagram shows a process in which the tunneling electron acquires energy from the previously excited dot. These processes are possible everywhere in the diamond.

From this rate we obtain the contribution of elastic cotunneling to the current from

$$I_{\rm el} = eW_{DS}.$$

The above considerations are valid if no excited quantum dot states provide additional elastic cotunneling channels. For the case that the mean single-particle level spacing Δ in quantum dots is very small, the conductance was calculated in Averin and Nazarov, 1990, to be

$$G_{\rm el} = \frac{\hbar G_{\rm S} G_{\rm D} \Delta}{4\pi e^2} \left(\frac{1}{E_{\rm e}} + \frac{1}{E_{\rm h}} \right).$$

Here, $E_{\rm e} = \mu_{N+1} - E_{\rm F}$ is the separation of the next unoccupied $(N+1)$-electron ground state from the Fermi energy in the contacts, whereas $E_{\rm h} = E_{\rm F} - \mu_N$. The energy scale Δ is the mean single-particle level spacing in the quantum dot for a given spin orientation. For a two-dimensional dot it can be estimated via $\Delta \approx 2\pi\hbar^2/m^\star A$, where A is the quantum dot area. The conductances $G_{\rm S}$ and $G_{\rm D}$ are the conductances of the tunneling barriers connecting the dot to the source and drain contacts. The contribution of elastic cotunneling will therefore be stronger the larger the coupling of the dot is to source and drain.

Inelastic cotunneling Inelastic cotunneling (process 6, left-hand diagram in Fig. 18.37) is relevant at finite bias voltages. The final state of the quantum dot is higher by the excitation energy Δ than the initial state. This energy difference is supplied by the bias voltage. Therefore the process sets in, if $eV_{\rm SD} \geq \Delta$. Beyond this source–drain voltage, the current increases linearly with the applied voltage, i.e., $dI/dV_{\rm SD} \equiv G_{\rm inel}$ is constant. Therefore there is a step (the inelastic onset) when one measures $dI/dV_{\rm SD}$ as a function of $V_{\rm SD}$.

The measurement in Fig. 18.36(b) shows such steps in the differential conductance within the diamonds (arrows) which are a result of inelastic cotunneling processes. At the diamond boundaries each inelastic cotunneling onset meets a line of an excited state existing outside the diamonds. This is exactly the excited state that plays a role for the inelastic cotunneling transition.

The transition rates in the regime of inelastic cotunneling are calculated, like those for elastic cotunneling, using Fermi's golden rule. The amplitude of the tunneling process is

$$t_{\tau\ell} = \frac{t^\star_{\ell d} t_{\tau d'}}{\epsilon_\tau - \mu_N} + \frac{t^\star_{\ell d'} t_{\tau d}}{\mu_{N+1} - \epsilon_\ell},$$

where ℓ labels a state in the source contact and τ a state in the drain. The states d and d' label the two quantum dot states involved. This is the superposition of the two processes depicted in Fig. 18.39. The total tunneling rate is then given by

$$W_{\rm D\leftarrow S} = \frac{2\pi}{\hbar} \sum_{\ell\tau} |t_{\tau\ell}|^2 \, \delta(\epsilon_\tau - \epsilon_\ell + \Delta) f_{\rm S}(\epsilon_\ell)[1 - f_{\rm D}(\epsilon_\tau)].$$

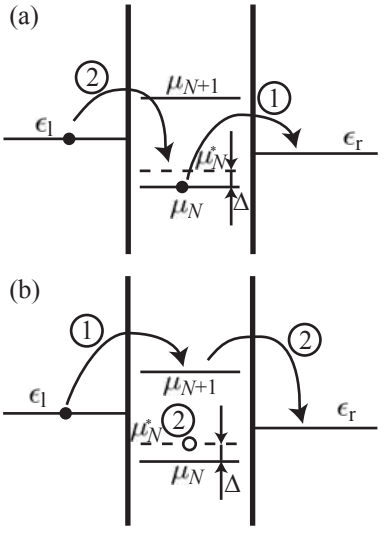


Fig. 18.39 Two processes contributing to inelastic cotunneling.

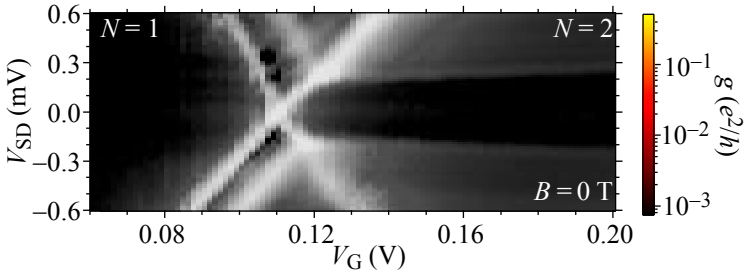


Fig. 18.40 Cotunneling through the excited triplet state of quantum dot helium. (Reprinted with permission from Zumbuhl *et al.*, 2004. Copyright 2004 by the American Physical Society.)

This tunneling rate will enter a rate equation used to calculate the occupation of the dot states. This is the case because the inelastic process populates, e.g., excited dot states, and therefore leads to a change of the occupation statistics of the dot. In addition to the inelastic cotunneling rate, decay rates also have to be taken into account that describe the decay of the excited state into the ground state by the emission of energy. (An example of such a calculation can be found in Wegewijs and Nazarov, 2001.)

The contribution of inelastic cotunneling to the conductance (i.e., the height of the cotunneling onset) has been calculated by Averin and Nazarov for the case where the single-particle level spacing Δ in the dot is much smaller than the charging energy. Their result is

$$G_{\mathrm{inel}} = \frac{\hbar G_{\mathrm{S}} G_{\mathrm{D}} \pi}{3e^2} (k_{\mathrm{B}}T)^2 \left(\frac{1}{E_{\mathrm{e}}} + \frac{1}{E_{\mathrm{h}}} \right)^2 .$$

The broadening of the step depends either on temperature, or on the intrinsic life time Γ_{inel} of the excited dot state. If $\Gamma_{\mathrm{inel}} \gg k_{\mathrm{B}}T$, the step has a width Γ_{inel}. If $\Gamma_{\mathrm{inel}} \ll k_{\mathrm{B}}T$, then the thermal broadening dominates and the step has the shape (Lambe and Jaklevic, 1968; Kogan *et al.*, 2004)

$$\frac{dI}{dV_{\mathrm{SD}}} = G_{\mathrm{el}} + G_{\mathrm{inel}} \left[F\left(\frac{eV_{\mathrm{SD}} + \Delta}{k_{\mathrm{B}}T} \right) + F\left(-\frac{eV_{\mathrm{SD}} - \Delta}{k_{\mathrm{B}}T} \right) \right],$$

where

$$F(x) = \frac{1 + (x-1)e^x}{(e^x - 1)^2}.$$

It turns out that the excited dot state reached by cotunneling has typically a larger life time than the corresponding excited state outside the Coulomb blockaded diamond. This finding can be exploited for doing excited state spectroscopy with enhanced precision (Franceschi *et al.*, 2001). The method was, for example, used in Zumbuhl *et al.*, 2004, for measuring the singlet–triplet transition of quantum dot helium in a magnetic field. Figure 18.40 shows the differential conductance of a quantum dot with one ($N = 1$) or two ($N = 2$) electrons. The excited triplet state can be seen outside the diamond. Within the $N = 2$ diamond there is a very clear cotunneling onset in the direction of V_{SD} related to the triplet state.

18.3.4 Tunneling with spin-flip: the Kondo effect in quantum dots

The Kondo effect has been known for a long time from measurements of the conductance of metals with magnetic impurities. At low temperatures particular metals follow different scenarios (see Fig. 18.41). While the resistance of normal metals without magnetic impurities tends towards a residual resistance given by the concentration of nonmagnetic impurities, the resistance of superconducting materials makes a very sharp transition to zero at the critical temperature. The resistance of metals with magnetic impurities, however, shows a logarithmic increase of the resistance as the temperature is lowered.

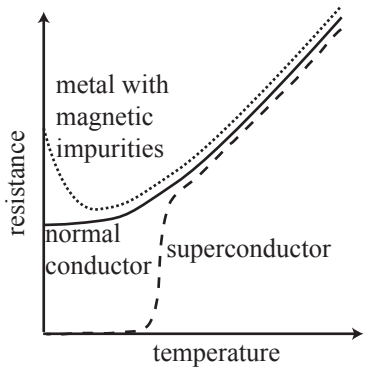


Fig. 18.41 Typical behavior of the resistance of a superconducting metal, a normal metal, and a metal with magnetic impurities as a function of temperature.

This increase in due to the Kondo effect. Conduction electrons scatter at the magnetic impurities and suffer (like the impurities) a spin-flip. It turns out that the spin-flip scattering cannot be treated in perturbation theory, but certain (Feynman) diagrams have to be summed up to infinite order. This effectively leads to a coherent screening of the localized impurity spin by the spin of the surrounding conduction band electrons (Kondo cloud). Crucial for this effect is the exchange interaction between the localized magnetic moment and the conduction band electrons. The correlated many-particle state of conduction band electrons and impurity spin is effectively a spin singlet which has an energy which is lowered by the energy scale $k_{\mathrm{B}}T_{\mathrm{K}}$. The temperature T_{K} is called the Kondo temperature. The effect is named after J. Kondo who developed the theory to explain the anomalous temperature-dependent resistance (Kondo, 1964; Kondo, 1969). The starting point of this theory is the so-called Anderson impurity hamiltonian.

The Kondo effect in quantum dots differs from the one in metals, because in quantum dots the important processes are those in which electrons are transmitted through the dot while suffering a spin-flip, whereas in metals it is spin-flip scattering that is important. The Kondo effect was predicted for tunneling through localized states at the end of the 1980s (Jones *et al.*, 1988; Glazman and Raikh, 1988; Ng and Lee, 1988).

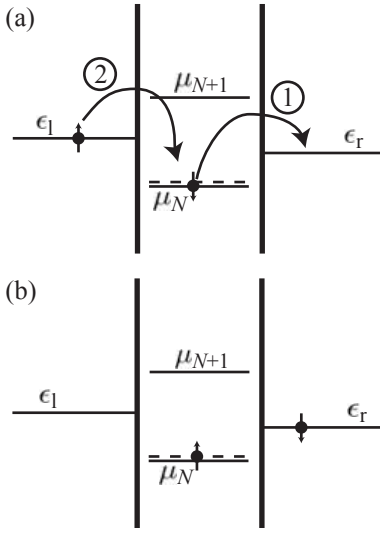


Fig. 18.42 Lowest order cotunneling process contributing to the Kondo effect. (a) Initial state and cotunneling processes. (b) Final state after the correlated tunneling event.

The Kondo effect occurs in quantum dots if a spin-degenerate state exists which is occupied with a single unpaired electron, as schematically shown in Fig. 18.42. For the discussion we suppose that the unpaired electron initially has spin down. The lowest order cotunneling process contributing to the Kondo effect is shown in Fig. 18.42(a). First, the electron tunnels from the dot into the drain contact gaining the energy $\epsilon_\tau - \mu_N$. In a correlated step an electron with the opposite spin tunnels from the source contact into the quantum dot losing the energy $\epsilon_\ell - \mu_N$. If $\epsilon_\tau = \epsilon_\ell$, then the correlated process conserves energy, but the quantum dot has reversed the direction of its unpaired spin [Fig. 18.42(b)]. Summing all processes of even higher order, the tunneling density of states develops a sharp resonance at the Fermi energy of source and drain, as shown in Fig. 18.43.

This peak in the tunneling density of states at the Fermi energy leads

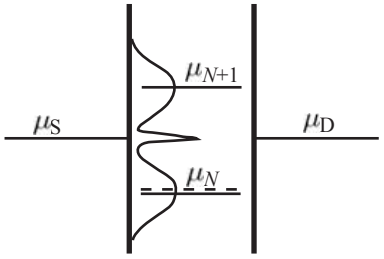


Fig. 18.43 Tunneling density of states for the Kondo effect exhibiting a sharp peak at the Fermi energy of the source and drain reservoirs.

to an enhanced current in the Coulomb blockade at zero source–drain voltage. The effect becomes visible in the diamonds as a zero-bias anomaly. The first experiments on quantum dots where this effect was demonstrated were published in Goldhaber-Gordon *et al.*, 1998. Figure 18.44(a) shows a corresponding measurement by Schmid *et al.*, 2000. The zero-bias anomaly is the dark stripe of enhanced conductance at zero source–drain voltage.

The characteristic temperature scale of the Kondo effect, the Kondo temperature T_K is given by

$$k_B T_K = \frac{1}{2}\sqrt{\Gamma U}\, e^{\pi\epsilon_0(\epsilon_0+U)/\Gamma U}. \tag{18.58}$$

Here $\epsilon_0 = \mu_{S/D} - \mu_N$ is the energetic separation of the Fermi energy in the contacts, and the spin-degenerate N-electron transition, $U = \mu_{N+1} - \mu_N$ is the addition energy for the next electron, and $\Gamma = \Gamma_S + \Gamma_D$ is the coupling of the dot to the leads. The Kondo effect can only be observed for temperatures $T < T_K$. The temperature dependence of the effect is shown in Fig. 18.44(b) and (c). The amplitude of the Kondo peak at zero source-drain voltage decreases dramatically with increasing temperature. Equation (18.58) shows that the Kondo temperature is higher the larger the tunneling coupling Γ of the dot to the leads. Therefore the Kondo effect is mainly observed in the strong coupling regime in which the Coulomb blockade diamonds are already strongly smeared. Furthermore, the Kondo temperature increases as one of the two ground state transitions μ_N and μ_{N+1} is tuned closer to the Fermi energies in the contacts, and as a consequence the amplitude of the Kondo peak increases [cf., Fig. 18.45(b)]. A further condition for the observation of the Kondo effect is that the mean spacing Δ between excited quantum dot states is larger than the tunneling coupling Γ. This condition requires dots that are small in size.

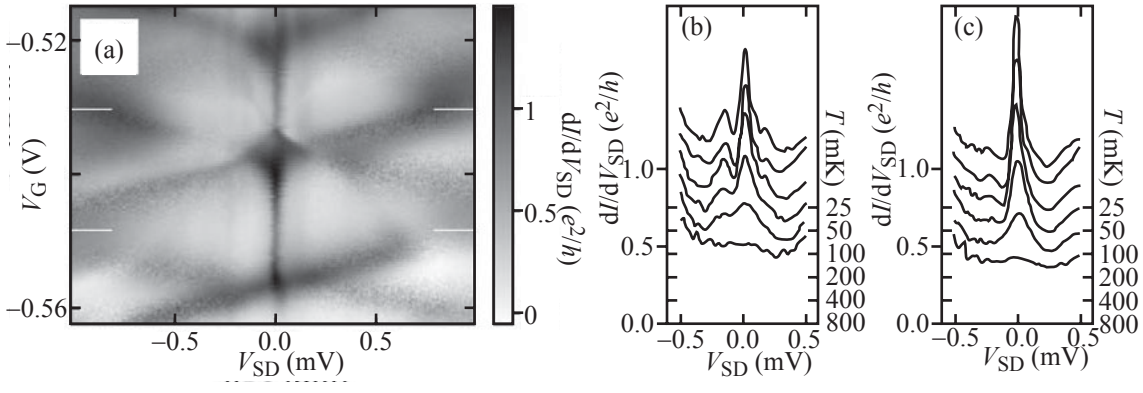


Fig. 18.44 (a) Differential conductance dI/dV_{SD} of a quantum dot as a function of the source–drain voltage V_{SD} and the gate voltage V_G. The Kondo resonances can be seen as dark stripes of enhanced differential conductance at zero source–drain voltage. (b) Cross sections through the diamonds in the middle between the two upper conductance peaks visible in (a) for different temperatures. (c) The same for the valley between the two lower conductance peaks. (Reprinted with permission from Schmid *et al.*, 2000. Copyright 2000 by the American Physical Society.)

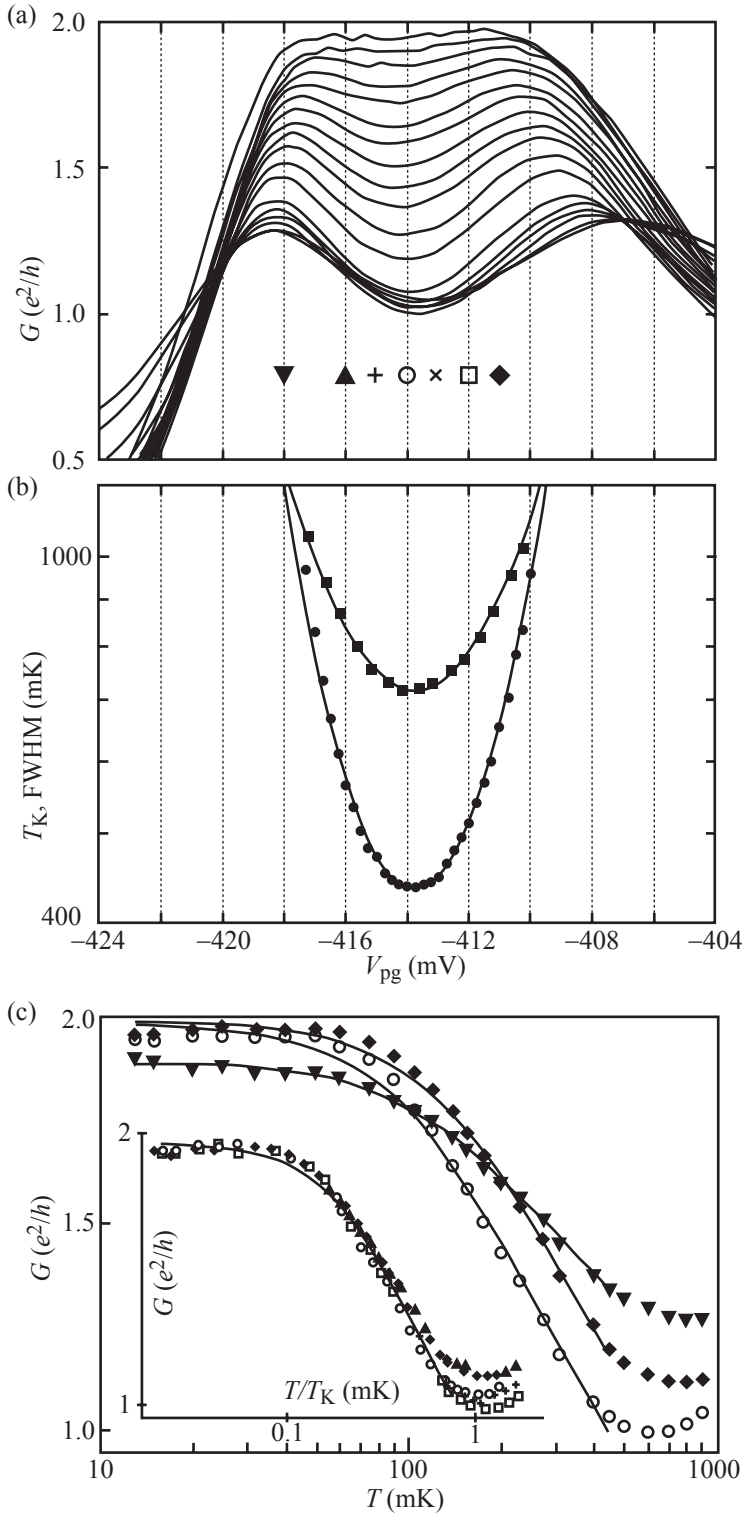


Fig. 18.45 Temperature dependence of the Kondo peak. (a) Gate voltage dependence of the conductance at zero source–drain voltage. (b) Kondo temperature as a function of gate voltage. (c) Scaling of the peak amplitude according to eq. (18.59) (van der Wiel *et al.*, 2000).

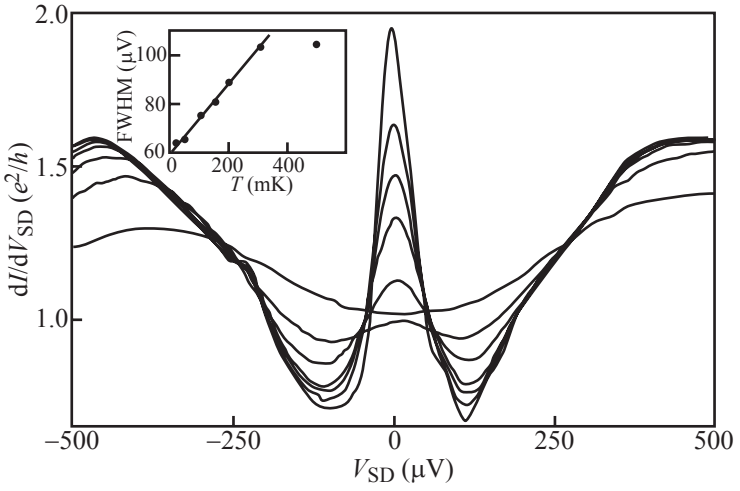


Fig. 18.46 Temperature dependence of the Kondo peak amplitude. The maximum amplitude (unitary limit) is almost reached (van der Wiel *et al.*, 2000).

The temperature dependence of the Kondo peak follows the (empirical) scaling law (van der Wiel *et al.*, 2000)

$$\frac{G(T)}{G_0} = \left[1 + \left(2^{1/s} - 1\right)\left(\frac{T}{T_\mathrm{K}}\right)^2\right]^{-s}. \tag{18.59}$$

Here, $s \approx 0.2$ for a system with spin 1/2 and $G_0 = 2e^2/h$. According to this law, the maximum height of the Kondo peak is G_0. This limit was indeed verified experimentally (van der Wiel *et al.*, 2000), as shown in Fig. 18.46. Another representation of the same data in Fig. 18.45(c) shows scaling according to eq. (18.59). The dependence of the Kondo temperature on the gate voltage mentioned above has been plotted in Fig. 18.45(b).

The Kondo effect has given rise to a vast amount of research, both experimentally and theoretically. In particular, a variant with integer spin has been found (Sasaki *et al.*, 2000; Pustilnik *et al.*, 2001), and there exists an orbital version of the Kondo effect that does not involve spin at all.

Further reading

- Books about quantum dots: Jacak *et al.* 1998; Chakraborty 1999.
- Books containing quantum dot physics: Grabert and Devoret 1992; Ferry 1998; Datta 1997; Ando *et al.* 1998; Heinzel 2007.
- Most important review of electronic transport: Kouwenhoven *et al.* 1997.
- Further reviews: Kastner 1992; von Klitzing 1996; Reimann 2002.
- Reviews of random matrix theory: Beenakker 1997; Alhassid 2000; Aleiner *et al.* 2002.
- Paper: Kouwenhoven *et al.* 2001.

Exercises

(18.1) A single electron is trapped in a potential box of typical size 200 nm. Estimate the electrostatic charging energy required to add a second electron to the well. Compare this energy with the typical single-particle level splitting in the box. How do your results depend on the relative dielectric constant of the material in which the box is realized?

(18.2) Consider the following three quantum dot systems:

(a) A lateral quantum dot fabricated from a two-dimensional electron gas in a Ga[Al]As heterostructure [see inset of Fig. 18.1]. The dot with electronic size 300 nm × 300 nm is defined using metallic gate electrodes.

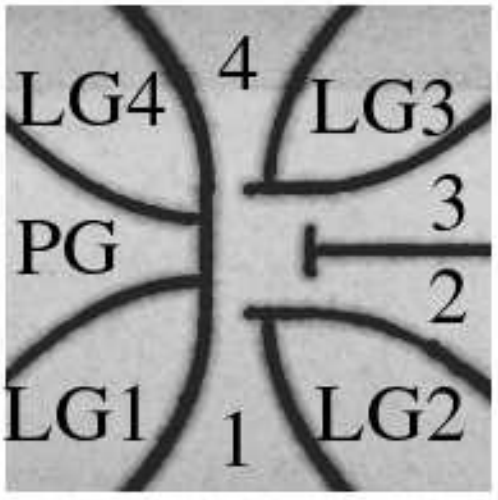
(b) A vertical quantum dot based on a 10 nm wide AlGaAs/InGaAs/AlGaAs quantum well structure with highly doped contacts above and below the quantum well. The quantum dot is defined by etching vertical cylinder structures with a diameter of about 0.5 µm. The electronic diameter of the quantum dot is about 100 nm. Such a dot is shown in Fig. 18.18.

(c) A self-assembled InAs quantum dot with a lateral diameter of 15 nm and a height of 3 nm.

(d) A quantum dot formed between the two metallic contacts of a carbon nanotube. The two contacts allow tunneling coupling into the tube, and their separation is 1 µm. The diameter of the nanotube is 2 nm.

Estimate for these four systems the charging energy and the single-particle level spacing. Compare these two energy scales for each system, and compare energy scales between systems. Discuss which of these four systems could allow the observation of Coulomb blockade at room temperature. Hint: The effective mass in the nanotube can be taken to be $m^\star = 0.06m$.

(18.3) The figure below shows a scanning force microscope image of the surface of a Ga[Al]As heterostructure in which a four-terminal quantum dot has been defined by local anodic oxidation (Leturcq *et al.*, 2004). The quantum dot is connected to external ohmic contacts. For simplicity we pinch off the connection to terminal 4 such that only three terminals remain relevant. We describe the strength of the coupling between the dot and the three leads by tunneling rates $\Gamma_i \ll k_\mathrm{B}T$ $(i = 1, 2, 3)$.



(a) Set up the rate equations for the occupation of a single quantum dot level (neglect any other levels).

(b) Convince yourself that the system is conveniently described in the linear response regime by a 3×3 conductance matrix.

(c) Relate the occupation probabilities from the rate equations with the three currents I_1, I_2, and I_3.

(d) Solve the rate equations and calculate the elements of the conductance matrix in linear response.

(e) Under what condition would you observe conductance resonances in the three currents? Do the resonances occur in all currents at the same energy of the quantum dot level?

(f) Which parameter determines the width of the resonances. What is their functional form when measured as a function of the plunger gate voltage PG?

This page intentionally left blank

Coupled quantum dots

19

Individual quantum dots can be mutually coupled by making their separation very small. Usually two contributions to the coupling have to be considered: On the one hand, the electrostatic interaction between the electrons of neighboring dots leads to a mutual influence on the energy spectra. This type of coupling is often called capacitive coupling. On the other hand, tunneling coupling may arise between neighboring quantum dots leading to a splitting of resonant energy levels.

Figure 19.1(a) shows the conductance measured on a double quantum dot structure of the type shown in Fig. 19.2(b). Conductance resonances show characteristic kinks, and they follow one of two characteristic slopes in the plane of the two plunger gate voltages. The corresponding sample depicted in Fig.19.1(b) is based on a shallow two-dimensional electron gas which is laterally patterned by AFM lithography. In the following we will discuss how the hexagon-shaped regions enclosed by conductance resonances and forming a honeycomb pattern come about, and what their meaning is.

Figure 19.2 shows schematically different arrangements of two quantum dots. They may be either connected in series (a), or they are connected in parallel (b, c) between a source and a drain contact. Plunger gates 2 and 3 allow us to control the number of electrons in the two quantum dots. As a consequence of the smallness of the structure, additional capacitances have to be considered beyond those indicated between the gates and the dots. For example, gate 2 will also act on dot 1, and gate 3 will tune dot 0.

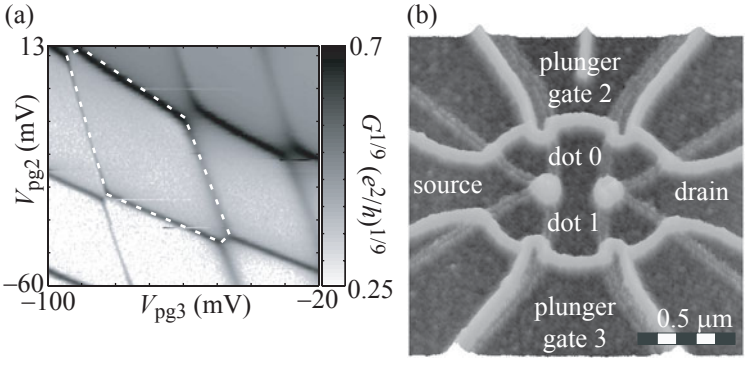


Fig. 19.1 (a) Conductance of the parallel double quantum dot system depicted in (b). The conductance resonances are measured in the plane of the two plunger gates and enclose hexagon-shaped regions, one of which is indicated with a white dashed line. The two quantum dots are weakly coupled by tunneling.

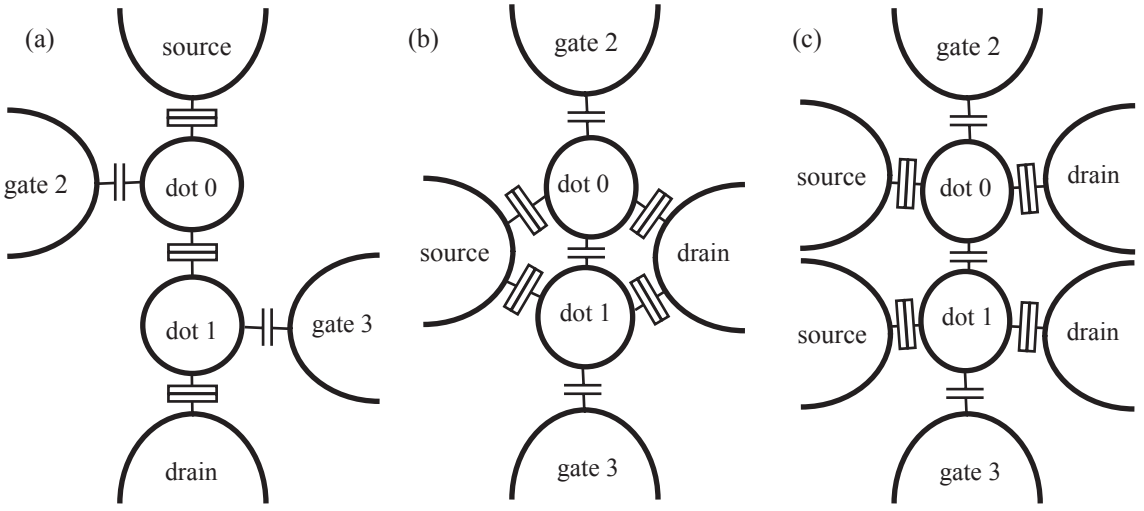


Fig. 19.2 Schematic arrangement of typical double quantum dot structures. (a) Two dots connected in series and placed between a source and a drain contact. (b) Two dots connected in parallel between the same pair of source and drain contacts. (c) Two dots connected in parallel with each dot having a separate pair of source and drain contacts.

19.1 Capacitance model

In order to obtain insight into the states of a double dot system, we consider only electrostatic coupling between the quantum dots. The realization of a system which is of type (c) in Fig. 19.2 is shown in Fig. 19.3. Within the capacitance model we express the coupling between the dots and the metallic electrodes of the system by a capacitance matrix. The general expression for the charge Q_i on the ith electrode is given by eq. (18.11) which we repeat here for convenience:

$$Q_i = \sum_{j=0}^{n} C_{ij}\phi_j + Q_i^{(0)}.$$

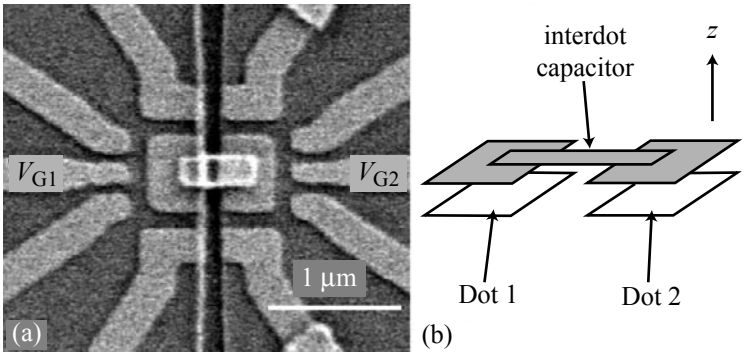


Fig. 19.3 Double dot system with purely capacitive coupling between the dots. (a) Scanning electron microscope image of the structure. The gate labeled V_{G1} is the plunger gate of the left dot, that labeled V_{G2} is the plunger gate of the right dot. An etched trench which does not allow tunneling separates the two dots. A metallic bridge from one dot to the other across the trench strengthens the capacitive coupling. (b) Schematic illustration showing how the metallic bridge enhances the capacitive coupling between the dots. (Reprinted from Chan *et al.*, 2003 with permission from Elsevier.)

For the two coupled quantum dots we use the indices $i = 0, 1$. The electrostatic potentials $\phi_{0,1}$ of the dots are unknown, but we consider the charges $Q_{0,1}$ to be known. Using the equations above we can express the dot potentials $\phi_{0,1}$ as a function of the other gate potentials and the charges $Q_{0,1}$. We introduce the abbreviation

$$\begin{aligned} A_0 &= Q_0 - Q_0^{(0)} - \sum_{j=2}^{n} C_{0j}\phi_j \\ A_1 &= Q_1 - Q_1^{(0)} - \sum_{j=2}^{n} C_{1j}\phi_j \end{aligned}$$

and find the relation

$$\begin{pmatrix} A_0 \\ A_1 \end{pmatrix} = \begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix} \begin{pmatrix} \phi_0 \\ \phi_1 \end{pmatrix},$$

which can be inverted to give

$$\begin{pmatrix} \phi_0(Q_0, Q_1) \\ \phi_1(Q_0, Q_1) \end{pmatrix} = \frac{1}{C_{00}C_{11} - C_{01}C_{10}} \begin{pmatrix} C_{11} & -C_{01} \\ -C_{10} & C_{00} \end{pmatrix} \begin{pmatrix} A_0 \\ A_1 \end{pmatrix}.$$

The total electrostatic energy of the double dot system is then calculated in complete analogy to the single dot case from

$$\begin{aligned} E_{\mathrm{elstat}}(N_0, N_1) = &\int_{Q_0^{(0)}}^{Q_0^{(0)} - eN_0} \phi_0(Q_0, Q_1 = 0) dQ_0 \\ &+ \int_{Q_1^{(0)}}^{Q_1^{(0)} - eN_1} \phi_1(Q_0 = Q_0^{(0)} - eN_0, Q_1) dQ_1. \end{aligned}$$

The result of the integration with the potentials given above is

$$\begin{aligned} E_{\mathrm{elstat}}(N_0, N_1) = &\frac{e^2 N_0^2}{2C_{\Sigma 0}} + \frac{e^2 N_1^2}{2C_{\Sigma 1}} + \frac{e^2 N_0 N_1}{\tilde{C}_{01}} \\ &- eN_0 \sum_{j=2}^{n} \alpha_{0j}\phi_j - eN_1 \sum_{j=2}^{n} \alpha_{1j}\phi_j, \end{aligned}$$

where

$$\begin{aligned} C_{\Sigma 0} &= C_{00}\left(1 - \frac{C_{10}C_{01}}{C_{00}C_{11}}\right) > 0, \\ C_{\Sigma 1} &= C_{11}\left(1 - \frac{C_{10}C_{01}}{C_{00}C_{11}}\right) > 0, \\ \tilde{C}_{01} &= \frac{C_{00}C_{11} - C_{01}C_{10}}{-C_{01}} > 0, \\ \alpha_{0j} &= \frac{C_{01}C_{1j} - C_{11}C_{0j}}{C_{00}C_{11} - C_{01}C_{10}} > 0, \\ \alpha_{1j} &= \frac{C_{10}C_{0j} - C_{00}C_{1j}}{C_{00}C_{11} - C_{01}C_{10}} > 0. \end{aligned}$$

In order to obtain the total energy of the double quantum dot system in our simple model, we add to the total electrostatic energy of the two quantum dots the quantization energies of the levels in the two dots and neglect tunneling coupling between them. We obtain

$$E(N_0,N_1)=\underbrace{\sum_{n=0}^{N_0}\epsilon_n^{(0)}+\frac{e^2N_0^2}{2C_{\Sigma 0}}-eN_0\sum_{j=2}^{n}\alpha_{0j}\phi_j}_{\text{Dot 0}}+\underbrace{\sum_{n=0}^{N_1}\epsilon_n^{(1)}+\frac{e^2N_1^2}{2C_{\Sigma 1}}-eN_1\sum_{j=2}^{n}\alpha_{1j}\phi_j}_{\text{Dot 1}}+\underbrace{\frac{e^2N_0N_1}{\tilde{C}_{01}}}_{\text{INT}}.$$

Comparing with the total energy of a single quantum dot in eq. (18.12) we are led to the following interpretation: the total energy of the coupled system is the sum of the energies of the individual dots plus an electrostatic coupling energy (INT) containing the product N_0N_1 and the mutual capacitance $\tilde{C}_{01}$.

As in the case of a single quantum dot we can ask for the energy required to add a single electron to dot 0 (dot 1) while keeping the charge in dot 1 (dot 0) constant at the value N_1 (N_0). We call this quantity the electrochemical potential of dot 0 (dot 1). It is given by

$$\begin{aligned}\mu_{N_0}^{(0)}(N_1) &= \epsilon_{N_0}^{(0)}+\frac{e^2}{C_{\Sigma 0}}\left(N_0-\frac{1}{2}\right)-e\sum_{j=2}^{n}\alpha_{0j}\phi_j+\frac{e^2}{\tilde{C}_{01}}N_1\\ \mu_{N_1}^{(0)}(N_0) &= \epsilon_{N_1}^{(1)}+\frac{e^2}{C_{\Sigma 1}}\left(N_1-\frac{1}{2}\right)-e\sum_{j=2}^{n}\alpha_{1j}\phi_j+\frac{e^2}{\tilde{C}_{01}}N_0.\end{aligned}$$

For example, if we do tunneling spectroscopy of the $\mu_{N_0}^{(0)}$, the energy of these levels depends on the charge state of dot 1. An additional electron in dot 1 shifts the whole addition spectrum in dot 0 up in energy. In turn, the same is true for the spectrum of dot 1.

A special situation arises when at certain gate voltages $\mu_{N_0}^{(0)}=\mu_{N_1}^{(1)}$. In this case, the two dots are in resonance, and an electron can be shifted from one dot to the other without energy cost. If the tunneling coupling between the dots is sufficiently strong, the two resonant levels are further split. This tunneling splitting will be discussed later.

Charge stability diagram. Now we investigate the details of the so-called charge stability diagram of the double dot system. In a typical experiment (cf., Fig. 19.3) the charge on each of the two quantum dots is tuned with a separate plunger gate while all other gate voltages stay constant. The two tuned plunger gate voltages—we call them ϕ_2 for dot 0 and ϕ_3 for dot 1 (see Fig. 19.2)—span a two-dimensional parameter plane. In each point of this plane, i.e., for each pair of parameters (ϕ_2,ϕ_3), there is a charge state (N_0,N_1) which is the ground state of

Table 19.1 Transitions from the stable region (N_0, N_1) to other charge states in the parameter plane, and the corresponding conditions.

transition	condition	equation
$(N_0, N_1) \to (N_0+1, N_1)$	$\mu_{\rm S/D} = \mu^{(0)}_{N_0+1}(N_1)$	(19.1)
$(N_0, N_1) \to (N_0, N_1+1)$	$\mu_{\rm S/D} = \mu^{(1)}_{N_1+1}(N_0)$	(19.2)
$(N_0, N_1) \to (N_0-1, N_1)$	$\mu_{\rm S/D} = \mu^{(0)}_{N_0}(N_1)$	(19.3)
$(N_0, N_1) \to (N_0, N_1-1)$	$\mu_{\rm S/D} = \mu^{(1)}_{N_1}(N_0)$	(19.4)
$(N_0, N_1) \to (N_0+1, N_1-1)$	$\mu^{(0)}_{N_0+1}(N_1) = \mu^{(1)}_{N_1}(N_0+1)$	(19.5)
$(N_0, N_1) \to (N_0-1, N_1+1)$	$\mu^{(0)}_{N_0}(N_1) = \mu^{(1)}_{N_1+1}(N_0-1)$	(19.6)

the double dot system. Usually, neighboring points in the parameter plane will belong to the same ground state. Regions in the parameter plane belonging to the same charge ground state (N_0, N_1) are called charge stability regions. We will now find out which geometrical shape the charge stability regions have in the parameter plane. The charge in the double dot system is stable whenever the electrochemical potentials $\mu_{\rm S} = \mu_{\rm D}$ in the source and drain leads are *not* resonant with one of the two electrochemical potentials in the double dot. Boundaries of the charge stability region (N_0, N_1) are therefore reached if one of the six conditions is fulfilled given in Table 19.1. The six conditions in the right column of the table result in six linear equations in the parameter plane. These six boundary lines enclose a hexagonal area within which the state (N_0, N_1) is stable. The six equations are

$$\Delta\phi_3 = -\frac{\alpha_{02}}{\alpha_{03}}\Delta\phi_2 + \frac{1}{e\alpha_{03}}\left[\epsilon^{(0)}_{N_0+1} - \epsilon^{(0)}_{N_0} + \frac{e^2}{C_{\Sigma 0}}\right] \tag{19.1}$$

$$\Delta\phi_3 = -\frac{\alpha_{12}}{\alpha_{13}}\Delta\phi_2 + \frac{1}{e\alpha_{13}}\left[\epsilon^{(1)}_{N_1+1} - \epsilon^{(1)}_{N_1} + \frac{e^2}{C_{\Sigma 1}}\right] \tag{19.2}$$

$$\Delta\phi_3 = -\frac{\alpha_{02}}{\alpha_{03}}\Delta\phi_2 \tag{19.3}$$

$$\Delta\phi_3 = -\frac{\alpha_{12}}{\alpha_{13}}\Delta\phi_2 \tag{19.4}$$

$$\Delta\phi_3 = \frac{\alpha_{12}-\alpha_{02}}{\alpha_{03}-\alpha_{13}}\Delta\phi_2 + \frac{1}{e(\alpha_{03}-\alpha_{13})}\left[\epsilon^{(0)}_{N_0+1} - \epsilon^{(0)}_{N_0} + \frac{e^2}{C_{\Sigma 0}} - \frac{e^2}{\tilde{C}_{01}}\right] \tag{19.5}$$

$$\Delta\phi_3 = \frac{\alpha_{12}-\alpha_{02}}{\alpha_{03}-\alpha_{13}}\Delta\phi_2 - \frac{1}{e(\alpha_{03}-\alpha_{13})}\left[\epsilon^{(1)}_{N_1+1} - \epsilon^{(1)}_{N_1} + \frac{e^2}{C_{\Sigma 1}} - \frac{e^2}{\tilde{C}_{01}}\right] \tag{19.6}$$

Here we measure the voltages $\Delta\phi_2$ and $\Delta\phi_3$ from the zero point indicated in Fig. 19.4. We find the following properties of the hexagon boundaries: the lines (19.1) and (19.3) are parallel, like the lines (19.2) and (19.4),

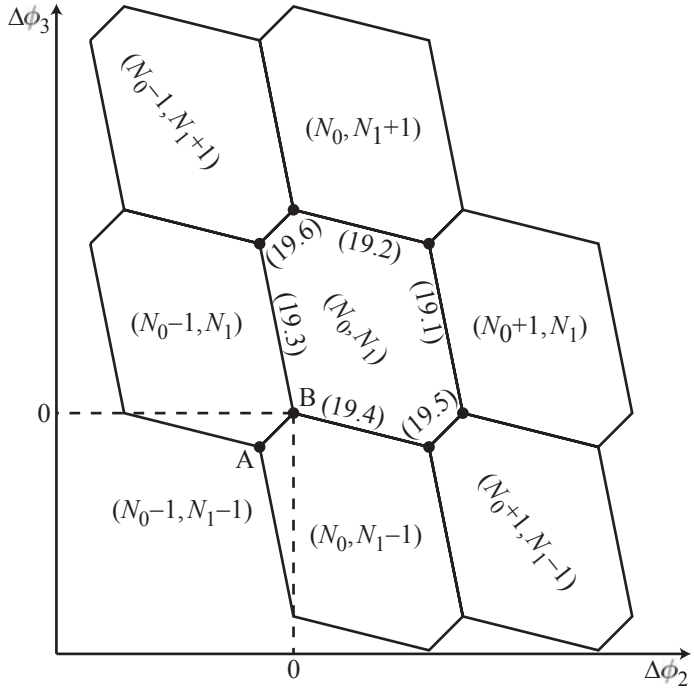


Fig. 19.4 Honeycomb pattern representing the charge stability diagram of a double quantum dot system in the capacitance model. Numbers (19.x) refer to the equation for the diamond boundary. Filled circles mark triple points, where three charge states coexist.

and (19.5) and (19.6). The slopes of the lines (19.1–19.4) are ratios of lever arms of the two gates acting on the two dots. These slopes are always negative. The slopes of the two lines (19.5) and (19.6) are usually positive. Figure 19.4 shows the typical shape of such hexagons and how they form a honeycomb pattern in the parameter plane. Such plots are called charge stability diagrams. In this diagram we can find diagonal lines from the top left to the bottom right along which the total charge $N_0 + N_1$ of the double dot system remains constant. Along these lines, the asymmetry (or polarization) between the two dots changes and is proportional to $N_0 - N_1$. If we follow diagonal lines from the bottom left to the top right, only the total number of electrons $N_0 + N_1$ is changed, but not the asymmetry $N_0 - N_1$. The six corners of each hexagon of stable charge are called *triple points*, because at these points, three charge states coexist. For example, in Fig. 19.4, at the triple point $\Delta\phi_2 = \Delta\phi_3 = 0$ (point B), the three charge states $(N_0{-}1, N_1)$, $(N_0, N_1{-}1)$, and (N_0, N_1) coexist. Triple points are particularly important for quantum dots connected in series because these are the only points in parameter space where an electron can be transported from source to drain resonantly.

Using the charge stability diagram of a double dot system we can now understand the results of the transport experiment made on the capacitively coupled quantum dots in Fig. 19.3 which are depicted in Fig. 19.5. In (a) the current through dot 1 is plotted in the parameter plane of the two plunger gates. The bright lines of enhanced current

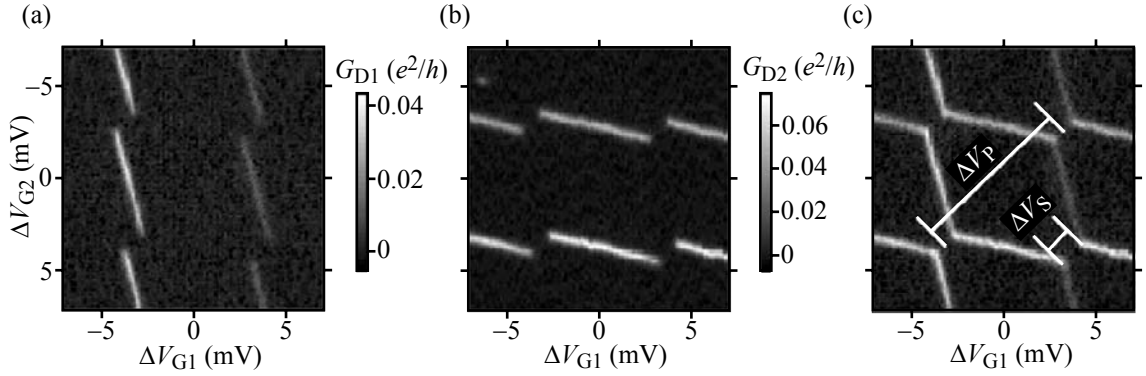


Fig. 19.5 Transport measurements through the two parallel quantum dots shown in Fig. 19.3 which are only capacitively coupled. (a) Conductance resonances in dot 1 as a function of the two plunger gate voltages. (b) Conductance resonances in dot 2 as a function of the same plunger gate voltages. (c) Superposition of the two measurements (a) and (b) resulting in the hexagon pattern of the capacitance model. (Reprinted from Chan *et al.*, 2003 with permission from Elsevier.)

are the conductance peaks of dot 1. Their finite slope is due to the fact that plunger gate 2 also couples to dot 1 and shifts the energy levels. Discontinuities in conductance resonance lines do always occur when an additional electron is charged onto dot 2. The mutual capacitive coupling $\tilde{C}_{01}$ of the two dots results in a shift of the spectrum of dot 1, whenever an additional charge appears on dot 2. The same analysis applies to Fig. 19.5(b), if the roles of dot 1 and dot 2 are interchanged. In (c) the superposition of the two measurements shown in (a) and (b) is shown. We can see the hexagon pattern characteristic for the double dot system.

19.2 Finite tunneling coupling

In the presence of a small tunneling coupling between the two dots, the charge stability diagram derived above changes only in the vicinity of the triple points where states of the two dots are degenerate. The tunneling coupling removes this degeneracy, and symmetric and antisymmetric states are formed.

In the following we describe the situation for the triple points labeled A and B in Fig. 19.4 assuming $N_0 = N_1 = 1$. In this case, at the triple point A the three charge states $(0,0)$, $(1,0)$, and $(0,1)$ coexist. The energy of the state $(0,0)$ with no electrons in the double dot system is taken to be $E_0 = 0$. For the one-electron situation, we regard the tunneling coupled double dot system as a single dot which can be called a *quantum dot molecule*, in which only two energy levels are relevant (one state of dot 1 and one of dot 2). The hamiltonian for one electron may then be taken to be

$$H = \begin{pmatrix} \epsilon_0(\phi_2, \phi_3) & \gamma_{01} \\ \gamma_{01}^\star & \epsilon_1(\phi_2, \phi_3) \end{pmatrix},$$

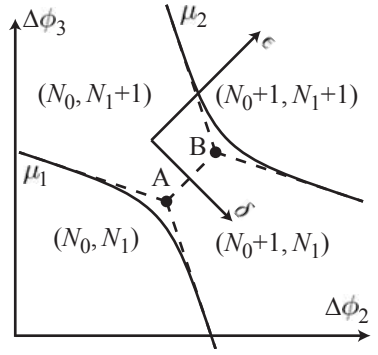


Fig. 19.6 Charge stability diagram of tunneling coupled quantum dots near the two triple points A and B. The tunneling coupling rounds the sharp kinks in the boundary lines of the hexagons. The figure contains the two arrows along which the total energy ϵ and the detuning δ change.

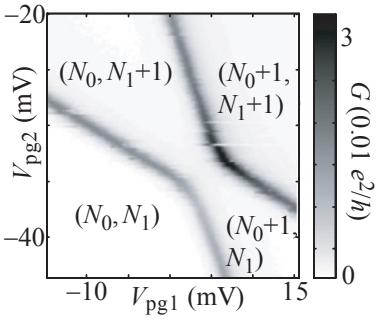


Fig. 19.7 Avoided crossing of two resonances in a double quantum dot measured in the configuration (b) of Fig. 19.2. The separation of the two triple points which is given by the capacitive coupling is further increased by the finite tunneling coupling between the dots.

where the tunneling coupling is described by the matrix element γ_{01}. The energies $\epsilon_{0,1}(\phi_2, \phi_3)$ are the energy levels of the system at zero tunneling coupling. They are tunable by the plunger gate voltages via the linear relations

$$\begin{aligned} \epsilon_0(\phi_2, \phi_3) &= \text{const.} - e\alpha_{02}\Delta\phi_2 - e\alpha_{03}\Delta\phi_3 \\ \epsilon_1(\phi_2, \phi_3) &= \text{const.} - e\alpha_{12}\Delta\phi_2 - e\alpha_{13}\Delta\phi_3. \end{aligned}$$

Diagonalizing the above hamiltonian gives the two one-electron energy eigenvalues

$$E_\pm = \frac{\epsilon_0 + \epsilon_1}{2} \pm \frac{1}{2}\sqrt{(\epsilon_0 - \epsilon_1)^2 + 4\left|\gamma_{01}\right|^2} = \epsilon \pm \frac{1}{2}\sqrt{\delta^2 + 4|\gamma_{01}|^2}.$$

The quantity $\delta = \epsilon_0 - \epsilon_1$ is called the *detuning* of the two quantum dot states, and $\epsilon = (\epsilon_1 + \epsilon_2)/2$ is the mean energy. These two quantities define a new rotated coordinate system in the plane of the two plunger gates as indicated in Fig. 19.6. The ground state energy E_1 of the one-electron quantum dot molecule is $E_1 = E_-$, because the state E_- is always lower in energy than E_+.

The energy of the two-electron system is given by

$$E_2 = 2\epsilon(\phi_2, \phi_3) + \frac{e^2}{\tilde{C}_{01}}.$$

The boundary lines of the hexagons near the triple points A and B are given by the electrochemical potentials of the quantum dot molecule, namely by

$$\mu_1 = E_1 - E_0 = \epsilon - \frac{1}{2}\sqrt{\delta^2 + 4\left|\gamma_{01}\right|^2}$$

near triple point A, and by

$$\mu_2 = E_2 - E_1 = \epsilon + \frac{1}{2}\sqrt{\delta^2 + 4\left|\gamma_{01}\right|^2} + \frac{e^2}{\tilde{C}_{01}}$$

near triple point B. We can see that for zero detuning δ between the two dots ($\epsilon_0 = \epsilon_1$), the separation between the two triple points reflects the sum of the capacitive coupling energy $e^2/\tilde{C}_{01}$ and the tunnel coupling splitting $2\gamma_{01}$. In the charge stability diagram this leads to the behavior depicted in Fig. 19.6. The sharp kinks observed at A and B without tunneling coupling become rounded. An experimental example where such a rounding of conductance resonances occurs near two triple points is shown in Fig. 19.7. In the extreme case of very strong coupling, the system no longer behaves like a double dot system, or a quantum dot molecule with electrons residing predominantly in one of the two dots, but like a single big quantum dot. The boundary lines between charge states $(N_0 + 1, N_1)$ and $(N_0, N_1 + 1)$ become increasingly irrelevant, because charges tend to be more and more shared between the two dots in large portions of the regions of stable charge. The boundary lines between states of distinct total charge of the system, however, still remain meaningful, but they stretch out to become almost straight lines running from the top left to the bottom right in the charge stability diagram. This will be further illustrated with experimental results in a later section.

19.3 Spin excitations in two-electron double dots

Spin states in double quantum dot systems have turned out to be of interest for quantum information processing to be discussed in a later chapter. We therefore discuss basic spin physics of one and two electrons in the double quantum dot system here. If there is only a single electron in one of the two dots, the spin behaves as in the single quantum dot case. Spin-up and spin-down states are degenerate at zero magnetic field. At finite magnetic field the Zeeman splitting lifts the spin degeneracy.

The situation becomes more interesting for the case of two electrons. There are three charge states compatible with two electrons, namely, $(2,0)$, $(0,2)$, and $(1,1)$. The two cases with two electrons in a single dot will favor singlet ground states which we label $(2,0)_{\mathrm{S}}$ and $(0,2)_{\mathrm{S}}$. The excited triplet states are in these two cases significantly higher in energy (for example, in lateral GaAs quantum dots the singlet–triplet splitting can be of the order of 1 meV or higher). The triplet states are degenerate at zero magnetic field, but a finite field lifts the degeneracy via the Zeeman effect and the three states T_+, T_0, and T_- can be distinguished.

In the case in which each dot holds one electron, the singlet–triplet splitting at zero magnetic field is usually negligibly small, as long as the two dots are only weakly coupled. At finite magnetic field, the T_- state is shifted down in energy and becomes the ground state below the T_0 and S states which stay degenerate. The T_+ state is even higher in energy at finite magnetic fields.

19.3.1 The effect of the tunneling coupling

Tunneling coupling between the dots leads to an avoided crossing of states with compatible spins as shown in Fig. 19.8. For example, at the boundary between the $(0,2)$ and the $(1,1)$ hexagon in the charge stability diagram, the state $(0,2)_{\mathrm{S}}$ is degenerate with $(1,1)_{\mathrm{S}}$ and $(1,1)_{\mathrm{T}}$ (this denotes all three triplet states degenerate at zero magnetic field). A finite tunneling coupling will lead to an avoided crossing of $(0,2)_{\mathrm{S}}$ and

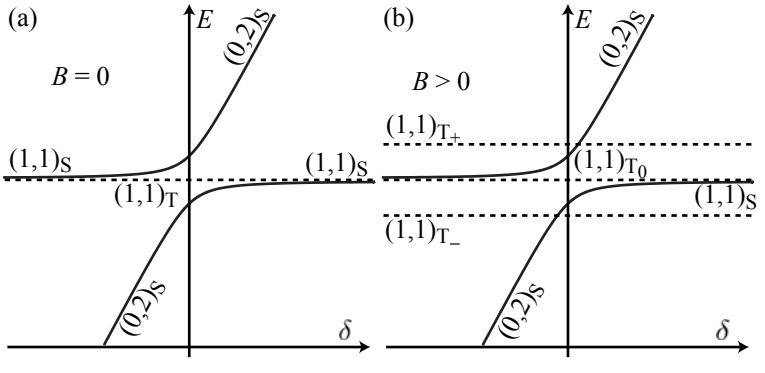


Fig. 19.8 (a) Energies of spin states close to the boundary between the $(1,1)$ and the $(0,2)$ hexagon at zero magnetic field as a function of detuning δ. Singlet states are represented by solid lines, triplet states are dashed. (b) The same for finite magnetic field, where the triplet states are Zeeman split.

Table 19.2 Magnetic properties of some nuclei important for semiconductors.

Isotope	natural abundance	I	μ/μ_N
^{12}C	98.9%	0	
^{13}C	1.1%	1/2	0.70
^{14}N	99.6%	1	0.40
^{15}N	0.4%	1/2	−0.28
^{27}Al	100%	5/2	3.64
^{31}P	100%	1/2	1.13
^{74}Ge	36.3%	0	
^{72}Ge	27.5%	0	
^{70}Ge	20.8%	0	
^{73}Ge	7.7%	9/2	−0.88
^{76}Ge	7.6%	0	
^{28}Si	92.2%	0	
^{29}Si	4.7%	1/2	−0.56
^{30}Si	3.1%	0	
^{69}Ga	60.1%	3/2	2.02
^{71}Ga	39.9%	3/2	2.56
^{75}As	100%	3/2	1.44
^{115}In	95.7%	9/2	5.54
^{113}In	4.3%	9/2	5.53
^{121}Sb	57.2%	5/2	3.36
^{123}Sb	42.8%	7/2	2.55

$(1,1)_S$ [Fig. 19.8(a)] because both spin states are singlet. The overlap integral of $(0,2)_S$ and $(1,1)_T$, however, is zero because the two spin states are orthogonal, and there is no avoided crossing. At finite magnetic fields the triplet states will be Zeeman split and the situation changes to that shown in Fig. 19.8(b).

19.3.2 The effect of the hyperfine interaction

The spatial probability distributions of electrons confined in quantum dots are extended over many lattice constants of the host crystal. Each electron can therefore interact with the spins of a large number (typically 10^6) of nuclei. For example in GaAs, the element Ga comes in two stable isotopes, ^{69}Ga and ^{71}Ga, which both have a nuclear spin 3/2, but different magnetic moments. In addition, As occurs naturally only as ^{75}As and has nuclear spin 3/2. The hyperfine interaction of an electron at position $\mathbf{r}$ with the nuclear spins at positions $\mathbf{R}_i$ is described by the Fermi contact hyperfine interaction hamiltonian

$$H_{\mathrm{HF}} = \frac{8\pi}{3}\frac{\mu_0}{4\pi} g_0 \mu_{\mathrm{B}} \sum_i \hbar \gamma_{\mathrm{N},i} \mathbf{I}_i \otimes \mathbf{S} \delta(\mathbf{r} - \mathbf{R}_i),$$

where μ_0 is the permeability of vacuum, $g_0 = 2.0023$ is the g-factor of the free electron, μ_{B} is Bohr's magneton, and $\mathbf{S}$ and $\mathbf{I}_i$ are the operators for the electron spin and nuclear spin, respectively. The quantity $\gamma_{\mathrm{N},i}$ is the gyromagnetic ratio of the nuclei. More detailed background about this hamiltonian is found in Slichter, 1963.

If we assume the electronic wave function to be a product of orbital and spin component, i.e., $|\psi(\mathbf{r})\rangle \otimes |\chi\rangle$, we can express the Fermi contact hyperfine interaction in terms of a pure spin hamiltonian as

$$H_{\mathrm{HF}} = \sum_i A_i \mathbf{I}_i \otimes \mathbf{S},$$

where

$$A_i = \frac{8\pi}{3}\frac{\mu_0}{4\pi} g_0 \mu_{\mathrm{B}} \hbar \gamma_{\mathrm{N},i} |\psi(\mathbf{R}_i)|^2$$

varies in space, i.e., with index i of the nucleus. The product of the spin operators can be written as

$$\mathbf{I}_i \otimes \mathbf{S} = \frac{1}{2}\left(S_+ \otimes I_- + S_- \otimes I_+\right) + S_z \otimes I_z$$

with $S_\pm = S_x \pm S_y$ and similar for $I_\pm$. The first term describes spin-transfer processes between the electronic system and the nucleus. For example, a spin flip in the electronic system leads to an inverse spin flip of the nucleus. The second expression is equivalent to an effective magnetic field in the z-direction caused by the nucleus which acts on the electron and causes an additional Zeeman splitting (on top of a Zeeman splitting due to an external magnetic field in the z-direction).

Indeed, the action of the nuclear spins can in good approximation be described by a classical magnetic field called the *Overhauser field*. An

electron in a state $\psi(\mathbf{r})$ experiences an effective Overhauser field (see, e.g., Zutic *et al.*, 2004)

$$\mathbf{B}_{\mathrm{N}} = \frac{1}{g^{\star}\mu_{\mathrm{B}}} \left\langle \sum_i A_i \mathbf{I}_i \right\rangle .$$

Spatial variations of this effective magnetic field lead to different precession velocities of electronic spins in two neighboring quantum dots, and therefore to an uncontrolled shift of their relative phase. This effect can be described by the effective spin hamiltonian

$$H = \frac{1}{2} g^{\star}\mu_{\mathrm{B}} (\mathbf{B}_{\mathrm{N}1}\sigma_1 + \mathbf{B}_{\mathrm{N}2}\sigma_2),$$

where $\mathbf{B}_{\mathrm{N}1}$ and $\mathbf{B}_{\mathrm{N}2}$ are the effective nuclear magnetic fields seen by the electrons in the two dots. If we write the operator part in brackets of the hamiltonian in matrix form using the basis functions $(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)/\sqrt{2}$ $((1,1)_{\mathrm{S}})$, $|\uparrow\uparrow\rangle$ $((1,1)_{\mathrm{T}_+})$, $(|\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle)/\sqrt{2}$ $((1,1)_{\mathrm{T}_0})$, $|\downarrow\downarrow\rangle$ $((1,1)_{\mathrm{T}_-})$, we find

$$\begin{pmatrix} 0 & \Delta_x - i\Delta_y & \Delta_z & -\Delta_x - i\Delta_y \\ \Delta_x + i\Delta_y & -\Sigma_z & \Sigma_x + i\Sigma_y & 0 \\ \Delta_z & \Sigma_x - i\Sigma_y & 0 & \Sigma_x + i\Sigma_y \\ -\Delta_x + i\Delta_y & 0 & \Sigma_x - i\Sigma_y & \Sigma_z \end{pmatrix},$$

where $\Delta_i = B_{1i} - B_{2i}$, and $\Sigma_i = (B_{1i} - B_{2i})/\sqrt{2}$ $(i = x, y, z)$. In Fig. 19.9 we plot a typical spectrum of this hamiltonian as a function of a magnetic field B applied in the z-direction and small components Δ_i. It can be seen that differences in the nuclear magnetic field between the two dots couple the spin singlet state to all three triplet states. If a finite external magnetic field much larger than the effective nuclear field is applied in the z-direction, the coupling between the triplet states $(1,1)_{\mathrm{T}_\pm} \equiv |\uparrow\uparrow\rangle, |\downarrow\downarrow\rangle$ and the singlet state will become irrelevant because these states differ appreciably in energy (Zeeman splitting). However, the $(1,1)_{\mathrm{T}_0}$ state will still mix with the singlet state. These two states will be split in energy by $g^{\star}\mu_{\mathrm{B}}\Delta B_z$ and combine to the new eigenstates $|\uparrow\downarrow\rangle$ and $|\downarrow\uparrow\rangle$.

The maximum effective nuclear magnetic field arises if all nuclear spins are aligned. In GaAs this results in a maximum nuclear field $B_{\max} = 5\,\mathrm{T}$. For randomly oriented spins, the effective nuclear magnetic field scales with $1/\sqrt{N}$, where N is the number of nuclear spins seen by the electron. Assuming $N \approx 10^6$, a typical value for the effective nuclear magnetic field in GaAs quantum dots is about 5 mT. The time evolution of the nuclear spin system turns out to be very slow compared to the electronic motion and the electronic spin precession. For more details on spins in few-electron quantum dot, Hanson *et al.*, 2007, is an excellent reference.

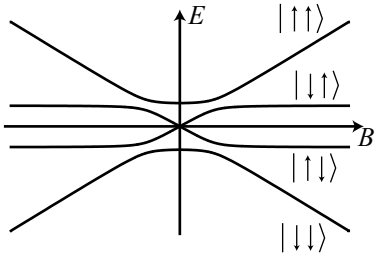


Fig. 19.9 Energy spectrum of a double quantum dot system with one electron in each dot as a function of external magnetic field B and small nuclear magnetic field differences Δ_i $(i = x, y, z)$.

19.4 Electron transport

19.4.1 Two quantum dots connected in parallel

The virtue of structures in which the two quantum dots are connected in parallel between a source and a drain contact is that the currents through the two parallel paths essentially add, and current can flow with either of the two, or both dots in resonance with the source and drain electrochemical potential. This leads to conductance measurements as a function of the two dots' plunger gates as shown in Fig. 19.1, or in Fig. 19.5. Similar structures have been investigated in Holleitner *et al.*, 2001, and Sigrist *et al.*, 2006.

Early experiments in which two quantum dots were connected in parallel with finite tunneling coupling between them were performed in Hofmann *et al.*, 1995. They observed and correctly interpreted the first hexagon patterned charge stability diagram. Quantum dots connected in parallel with mutual tunneling coupling can also be realized by vertical stacking. An implementation using a structure based on parallel quantum wells with separate in-plane contacts was realized in Wilhelm and Weis, 2000, an alternative with lateral coupling between the dots, but vertically stacked contacts was demonstrated in Hatano *et al.*, 2004.

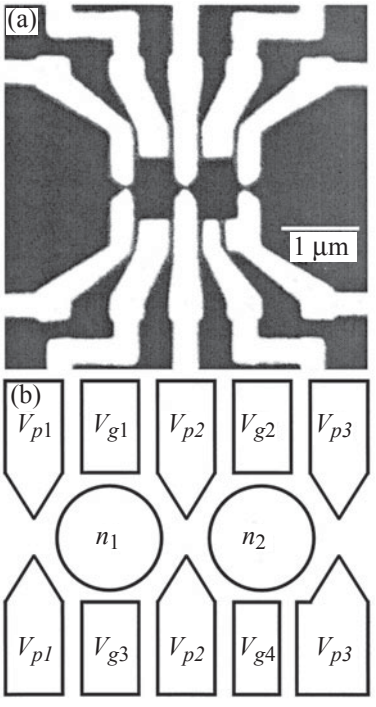


Fig. 19.10 Double dot structure fabricated by electron beam lithography on a Ga[Al]As heterostructure containing a two-dimensional electron gas. (a) Scanning electron micrograph of the double dot system. The white regions are metallic top gates. (b) Schematic drawing of the structure. The quantum dots are represented by circles (Livermore *et al.*, 1996).

19.4.2 Two quantum dots connected in series

Electron transport through quantum dots connected in series is different from both transport through single dots and transport through parallel dots. Figure 19.10 shows a structure in which two dots are connected in series. Electron transport in lowest order and at low source–drain bias is only possible at triple points of the charge stability diagram, where the electrochemical potentials of both dots are degenerate and aligned with the electrochemical potential of source and drain. This is schematically illustrated in Fig. 19.11. At the triple points (1) two energy levels in the dot are aligned with each other and with the electrochemical potential in the leads. An electron can be transferred from the source contact through both dots to the drain contact elastically. The situation is different at points (2) or (3), where the energy level in only one of the quantum dots is aligned with the electrochemical potential in one lead, whereas the other dot is in the Coulomb blockade. In this situation electron transfer from source to drain is exponentially suppressed in the case of weak tunneling coupling between the dots and between dots and leads. However, a cotunneling current may flow when the coupling between the dots and between the dots and the leads is increased.

Figure 19.12 shows the measured current in a serial double dot system in the parameter plane of the two plunger gates. The coupling between the two dots increases from (a) to (f) by increasing the gate voltage V_{p2}. For the weaker coupling (a), transport occurs only at triple points. However, the two neighboring triple points schematically shown in Fig. 19.11 are not resolved here, but pairs of triple points lead to the speckles of

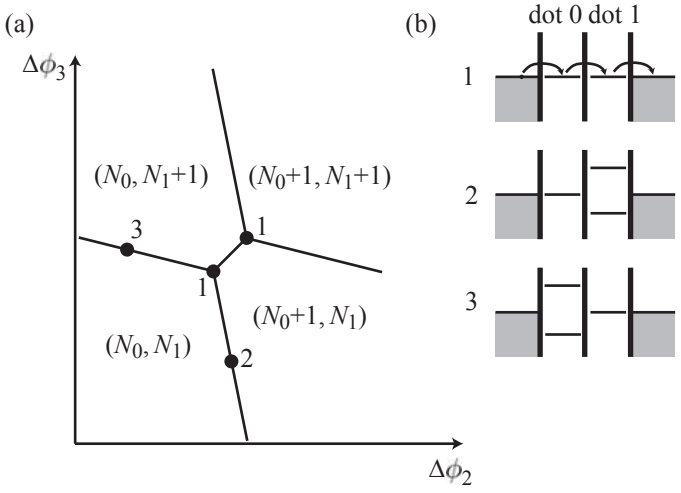


Fig. 19.11 (a) Part of a charge stability diagram. The energy level scheme of the double dot system is shown in (b) for the points marked with filled circles.

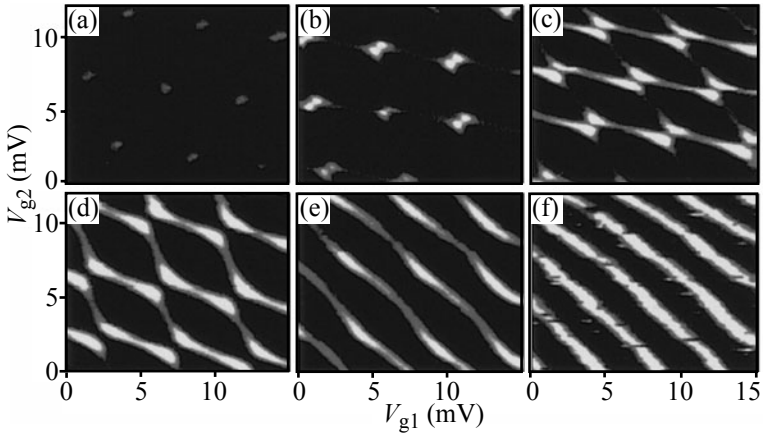


Fig. 19.12 Current through a double dot system with the dots connected in series between a source and a drain contact. The grayscale is logarithmic. The coupling between the dots is given by (a) $0.44e^2/h$, (b) $0.8e^2/h$, (c) $1.3e^2/h$, (d) $1.56e^2/h$, (e) $1.92e^2/h$ (Livermore *et al.*, 1996).

current flow. If the coupling between the dots is increased, the quantum dot molecule regime is reached. Correspondingly, the strict suppression of the current between the triple points is gradually lifted, and in (d) the hexagon pattern shown schematically in Fig. 19.4 and in measurements of parallel dots in Fig. 19.1 can be easily recognized. The kinks at the triple points are rounded by the coupling. In such a quantum dot molecule, electronic states are delocalized between the dots, and electrons are shared by both dots. In the extreme case shown in (f), the system appears to be a single Coulomb-blockaded quantum dot.

Spin blockade. When an electron tunnels through a potential barrier, its spin is usually conserved. In situations where the initial and the final state of the system have orthogonal spin configurations, the transition is forbidden and the tunneling current is strongly suppressed. This situation is called spin blockade.

Spin blockade can occur in single quantum dot systems with strong electron–electron interaction (Weinmann and Hausler, 1994; Weinmann *et al.*, 1995; Tanaka and Akera, 2006). For example, if the ground state of the N-electron system before tunneling has spin zero, and the ground state of the $(N+1)$-electron system after tunneling has spin 3/2, the addition of a single electron has to be accompanied by another spin-flip in the quantum dot. The tunneling transition may therefore be suppressed as a consequence of the fact that the initial spin 1/2 state (spin 0 of the dot electrons + spin 1/2 of the tunneling electron) is orthogonal to the final spin 3/2 state of the dot electrons. A different variant of spin-blockade physics in single quantum dots has been investigated in a regime where spin-polarized quantum Hall edge states exist in the leads and in the quantum dot (Imamura *et al.*, 1998; Ciorga *et al.*, 2000).

Here we will discuss the spin-blockade effect in double quantum dot systems (Ono *et al.*, 2002). The effect has been very clearly observed in double quantum dots with electron numbers below three (Johnson *et al.*, 2005), but it can also occur in dots with larger electron numbers, if closed shells are present. Figure 19.13 shows the charge stability diagram for a double quantum dot system with up to two electrons per dot. In order to appreciate the spin-blockade effect we will concentrate on the two pairs of triple points labeled (a) and (b) in the figure.

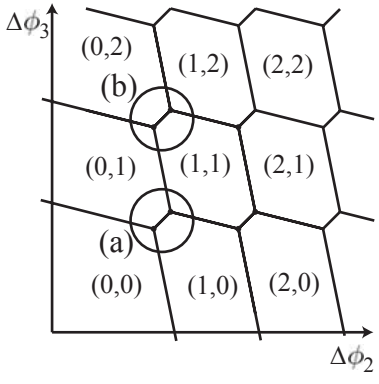


Fig. 19.13 Honeycomb pattern representing the charge stability diagram of a double quantum dot system for small electron numbers. Filled circles mark triple points, where three charge states coexist. At the encircled pair of triplet points labeled (a), single-electron or single-hole transport takes place at finite bias. At the encircled pair or triplet points labeled (b), one electron occupies the right dot already. Depending on the direction of the applied source–drain voltage, spin blockade can occur.

The spin-blockade effect is observed at finite source–drain voltage. Figure 19.14(a) shows the charge stability diagram close to the pair of triple points labeled (a) in Fig. 19.13 for a finite source–drain voltage applied to the double dot system. Stable charge states exist only in the unshaded regions as labeled in the figure. In the light gray regions the two adjacent charge states coexist (e.g., $(0,0)$ and $(0,1)$ in the leftmost light gray region). In the diamond-shaped dark gray regions, three charge states coexist (for example, (0,0), (0,1), (1,0) in the lower diamond). Such a triplet of states can be combined to form a transport cycle in which an electron is transferred from one contact to the other through the double quantum dot system [e.g., $(0,0) \rightarrow (1,0) \rightarrow (0,1) \rightarrow (0,0)$]. The two triangles forming such a diamond differ in the relative energy of

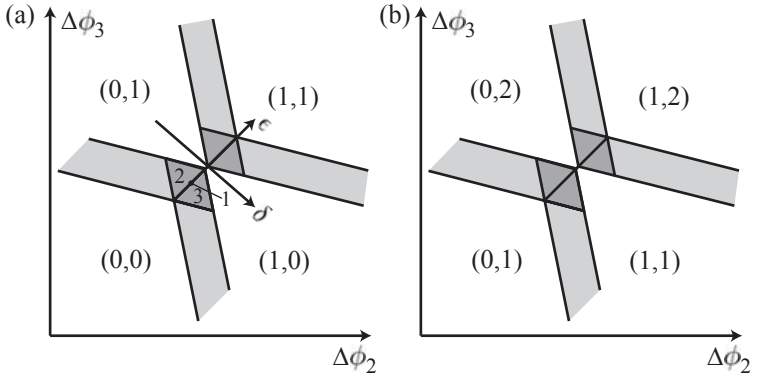


Fig. 19.14 Charge stability diagram of a double quantum dot system for small electron numbers under finite source–drain voltage. In the dark gray regions a full transport cycle is accessible. Fig. (a) corresponds to the pair of triple points labeled (a) in Fig. 19.13, whereas (b) corresponds to the pair of triple points labeled (b) in the same figure. The axis labeled δ in (a) is called the detuning axis between the states $(0,1)$ and $(1,0)$. The axis labeled ε is an energy axis. In this direction the detuning is unchanged, but the energies of the two states are tuned relative to the electrochemical potentials in source and drain.

the two states connected by a charge transfer from one dot to the other. For example, in the lower triangle, the state $(1,0)$ is lower in energy than the state $(0,1)$ as shown in scheme 3 of Fig. 19.15, whereas the opposite applies in the upper triangle (scheme 2 of Fig. 19.15). In case 2, (inelastic) transport can only take place if the electrochemical potential in the left (source) contact is higher than in the right (drain), whereas in case 3, the opposite is the case. Along the line where the two triangles touch, the two dot levels are aligned (which may lead to a level splitting due to finite tunneling coupling) as shown in Fig. 19.15, case 1, and elastic tunneling is possible. Figure 19.14(a) shows a coordinate system with axes labeled δ and ε. The δ-axis is called the detuning axis as the sum of the energies of the two dot states remains constant when the system state is changed along this direction, whereas the level separation (the detuning) between the two changes. The ε-axis is called the total energy axis because the detuning remains unchanged when the system state is changed along this direction, whereas the sum of the energies of the two dot states is altered.

For double quantum dots connected in series between source and drain, transport will only occur within one of the two dark gray triangles, depending of the polarity of the source–drain voltage. Figure 19.16(a) and (c) show the result of a measurement taken at this pair of triple points for forward (a) and reverse (b) source–drain voltage. In this measurement the source–drain voltage was chosen to be so large that the two triangles originating from the two triple points overlap. Along the line of zero detuning, where elastic tunneling is possible, the current is enhanced compared to the rest of the triangle. This is the scenario *without* the spin-blockade effect.

We now turn to the scenario depicted in Fig. 19.14(b) in which the right quantum dot already holds one electron. The ground state of the $(0,2)$ state will be a spin singlet state (S) $[(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)/\sqrt{2}]$ with the (0,2) spin-triplet excited states (T) being inaccessibly high in energy. However, the two spin states for the occupation numbers $(1,1)$ are es-

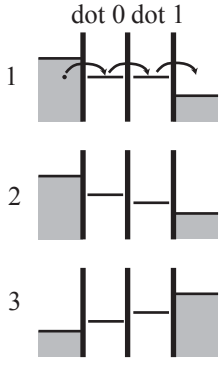


Fig. 19.15 Level alignment for electron transport in the finite bias triangles of Fig. 19.14. In scheme 1, the two levels are resonant, and a negative source–drain voltage is applied. In scheme 2, the two levels are negatively detuned ($\delta < 0$), and inelastic transport is possible at negative source–drain voltage. In scheme 3, the two levels are positively detuned, and transport is possible at positive source–drain voltage.

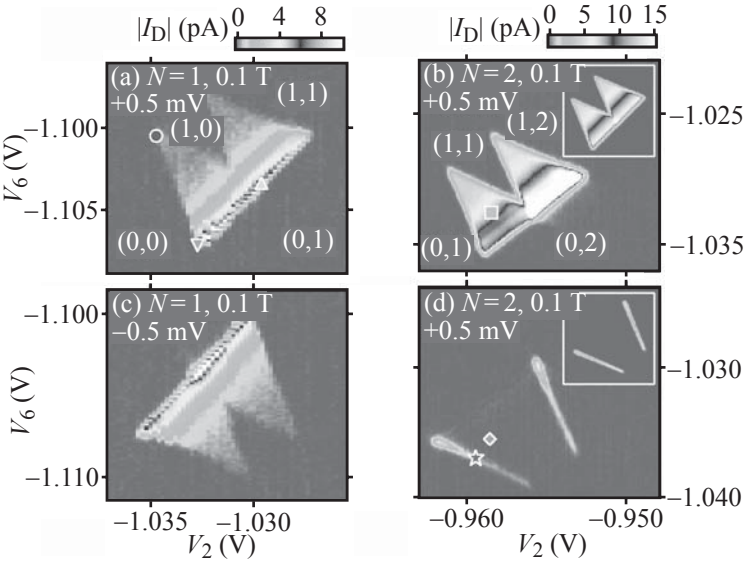


Fig. 19.16 Current through a double dot system at finite source–drain voltage. (a) and (c): Current through the double quantum dot system at positive (a) and negative (b) source–drain voltage for one excess electron tunneling through the empty system. (b) and (d): Current through the double quantum dot system at positive (b) and negative (d) source–drain voltage for one excess electron tunneling through the system already holding one electron in the second dot. In (d) the current is suppressed as a result of the spin-blockade effect. (Reprinted with permission from Johnson *et al.*, 2005. Copyright 2005 by the American Physical Society.)

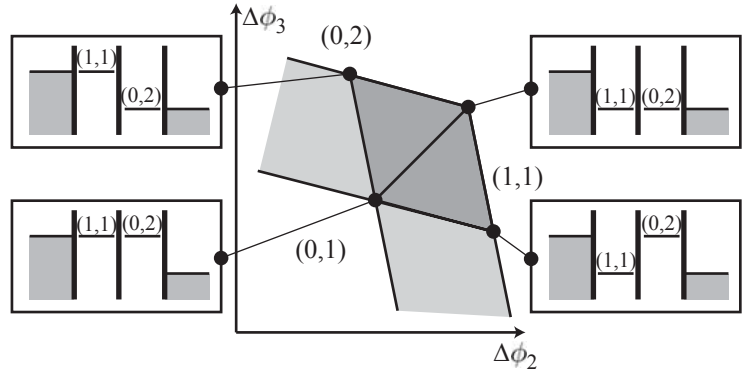


Fig. 19.17 Charge stability diagram and corresponding energy diagrams of a double quantum dot system for small electron numbers under finite source–drain voltage around a single triple-point. In the dark grey regions a full transport cycle is accessible.

sentially degenerate because the exchange interaction between the two electrons residing in different dots is suppressed by the central tunneling barrier. The energy level situation is the same as that depicted in Fig. 19.8(a). In the experiment, a finite magnetic field of 0.1 T defines a spin quantization axis without lifting this degeneracy appreciably. Figure 19.17 shows a magnification of the charge stability diagram around this particular triple point and the corresponding energy level schemes at the corners of the diamond shaped region.

At positive source–drain voltage (Fig. 19.18, scheme 2), the transport cycle is $(0,1) \rightarrow (0,2)_S \rightarrow (1,1)_S \rightarrow (0,1)$. It conserves the spin of the tunneling electron and current can flow, as observed in the measurement shown in Fig. 19.16(b). The situation is remarkably different for the opposite source–drain voltage polarity (scheme 1 in Fig. 19.18). Here, the tunneling electron can enter the left dot to form either the $(1,1)_S$ state, or the $(1,1)_T$ state, because they are at the same energy. In the first

case, the transport cycle will be $(0,1) \rightarrow (1,1)_S \rightarrow (0,2)_S \rightarrow (0,1)$ which conserves spin and therefore allows the electron to transfer. However, at some point an electron will enter the triplet state in the left dot, and the transport cycle $(0,1) \rightarrow (1,1)_T \rightarrow (0,2)_S \rightarrow (0,1)$ violates spin conservation in the transition $(1,1)_T \rightarrow (0,2)_S$. As a consequence, the electron cannot be transferred. It is stuck in the first dot because the finite applied source–drain voltage also prohibits tunneling back to the source contact. Remaining in the left dot, it inhibits any other electrons from entering the left dot owing to the Coulomb blockade effect. This is the spin-blockade situation which is experimentally observed in Fig. 19.16(d) as a suppression of the current in the two triangles at positive source–drain voltage. The spin-blockade is lifted in the experiment at the edges of the triangles, where the $(1,1)_T$ states are aligned with the electrochemical potential in the source contact because there an electron stuck in the triplet state can tunnel back to the lead and be replaced by an electron in the singlet state. Lifting the spin-blockade situation within the triangles requires a spin-flip process which can be mediated by hyperfine interaction with nuclear spins, or by spin–orbit interaction. The spin-blockade phenomenon can be exploited for the manipulation of individual spins in applications where the electron spin is used as the implementation of a qubit.

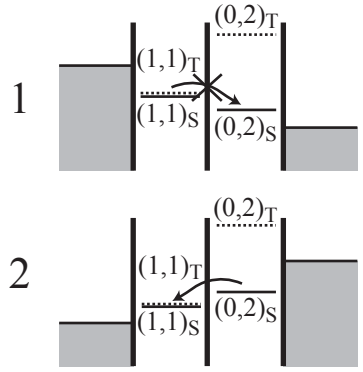


Fig. 19.18 Level alignment for electron transport in the finite bias triangles of Fig. 19.14(b). In scheme 1, the two levels are negatively detuned ($\delta < 0$) and inelastic transport is blocked by singlet–triplet spin blockade at negative source–drain voltage. In scheme 3, the two levels are positively detuned and transport is possible at positive source–drain voltage.

Further reading

- Review: van der Wiel *et al.* 2003.
- Review of hyperfine interaction in double quantum dots: Hanson *et al.* 2007.
- Papers: Livermore *et al.* 1996; Johnson *et al.* 2005.

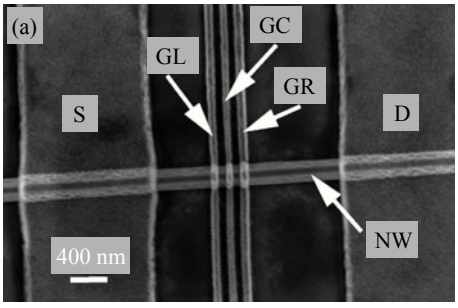
Exercises

(19.1) We consider a double quantum dot structure realized in an InAs ($\epsilon_r = 15$, electron density around $10^{18}\,\mathrm{cm}^{-3}$) nanowire (see the scanning electron microscope figure (a) below). In a purely classical description in which the influence of discrete quantum states is not taken into account, the double dot is modeled as a network of tunneling resistors and capacitors. The source–drain current I_{SD} has been measured at $V_{SD} = 140\,\mu\mathrm{V}$ and $T = 30\,\mathrm{mK}$ as a function of the left (V_{GL}) and right (V_{GR}) topgates. The double quantum dot behavior can be identified by the characteristic honeycomb pattern of the charge-stability diagram seen in figure (b).

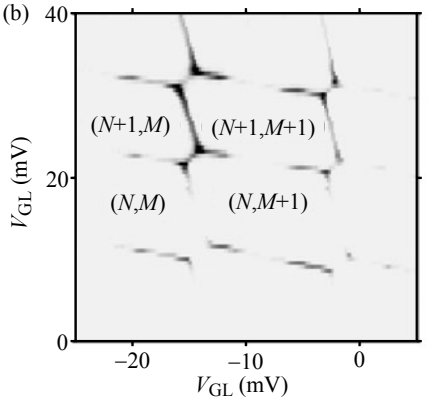
(a) Estimate the capacitances C_{GL} between left gate and left dot, and C_{GR} between right gate and right dot, and the total capacitances C_Σ for the two dots from the capacitive model and the measurement data.

(b) The lever arms of the gates were measured to be $\alpha_L = 0.46$, and $\alpha_R = 0.41$. With these values, it is possible to estimate the charging

energies of the left and the right quantum dot. How big are they?



(c) Assume the dots to be spherical, and estimate the radii of the two dots and the number of electrons in the dots.



Electronic noise in semiconductor nanostructures

20

20.1 Classification of noise

Certain physical processes within a conductor cause noise to appear in the current flow through the conductor. It turns out that some of these processes are inevitably related to the transport of electronic charges such that they cannot be avoided in principle. While the discontinuous emission of photons was investigated by Campbell in 1909 (Campbell, 1909*a*; Campbell, 1909*b*), first attempts to understand current noise were made in 1918 by Walter Schottky on vacuum tubes (Schottky, 1918). In order to find the origins of current noise in these tubes, all other sources of time-dependent current fluctuations in the remaining circuit had to be eliminated. This first experimental step poses the biggest experimental difficulty today, if one intends to investigate noise in semiconductor nanostructures. Schottky distinguished two types of noise:

- *thermal noise*. This contribution is often also called *Johnson–Nyquist noise*, after the experimentalist M.B. Johnson and the theoretician H. Nyquist, who studied thermal noise in detail (Johnson, 1927; Nyquist, 1928). Today we know that thermal noise is present in all electronic conductors at finite temperature. It does not require a finite mean current to flow through the conductor, but appears as soon as the two sides of the conductor are connected. It is therefore an equilibrium phenomenon.
- *shot noise*. Shot noise arises because the electronic charge is transported in quantized portions. Typically these portions have the size of the elementary charge $|e|$, but a few notable exceptions exist, e.g., Cooper-pairs in superconductors ($2|e|$), or fractionally charged quasiparticles in the fractional quantum Hall effect (e.g., $|e|/3$). Shot noise arises only if a finite mean current is driven through a conductor and is therefore a nonequilibrium phenomenon. Shot noise does not occur in all conductors. For example, in macroscopic metallic conductors shot noise is suppressed because the sample size exceeds the inelastic electronic mean free path by orders of magnitude. Individual segments of the material fluctuate

independently and the mean noise amplitude is strongly reduced by averaging. In mesoscopic samples, where the extent of the sample is smaller or comparable to the inelastic mean free path, shot noise is of importance. This is true for diffusive as well as for ballistic systems.

In addition to these two types of time-dependent current fluctuations, the so-called *random telegraph noise* frequently arises in mesoscopic systems. In the simplest case it manifests itself as a random switching of the current in time between two discrete values. The discrete jumps in the current can be explained as discrete jumps of the electrical resistance (or conductance) of the sample. It arises sometimes from thermally activated charging and discharging of single discrete charge traps in the vicinity of a current path. We have seen a random telegraph noise signal in Fig. 18.5 in the current of a quantum point contact that is able to detect the statistical charge fluctuations on a quantum dot in real time. We will come back to this example later in this chapter.

Another type of current noise is the so-called $1/f$-noise, or *flicker noise*. It got its name from the $1/f$-dependence of the noise current's power spectral density. It arises, for example, in commercially available carbon resistors, or in semiconductor transistors such as the Si-MOSFET. It is typically relevant at frequencies below about 10 kHz. The $1/f$-noise is also a nonequilibrium phenomenon because it requires a finite mean current to flow through the sample. A superposition of many individual noise sources leading to random telegraph noise on different time scales can be the origin of $1/f$-noise.

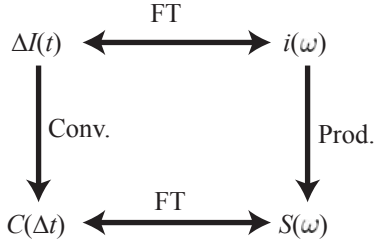


Fig. 20.1 Functions used to characterize electronic noise, and the relation between them.

20.2 Characterization of noise

When we considered the conductance or the resistance, we implicitly investigated the average current $\langle I \rangle$. Fluctuations of the current in time around this average

$$\Delta I(t) = I(t) - \langle I \rangle$$

are called the noise current, or simply the noise.

In order to characterize noise, a number of functions are used. These functions and their mutual relation will be introduced below. Figure 20.1 gives an overview where the different functions and their interrelations are represented graphically. On the top left is the noise current $\Delta I(t)$ introduced above. Convolution (Conv.) of the current with itself leads to the autocorrelation function $C(\Delta t)$. The Fourier transform (FT) of the autocorrelation function gives the spectral density $S(\omega)$. The latter can also be obtained from the squared modulus of the Fourier transformed noise current $i(\omega)$.

The autocorrelation function. The autocorrelation function is defined as

$$C(\Delta t) = \langle \Delta I(t) \Delta I(t + \Delta t) \rangle = \langle \Delta I(0) \Delta I(\Delta t) \rangle. \qquad (20.1)$$

It has the units [A^2], is a real-valued function, and does not depend on the time t explicitly, because $\Delta I(t)$ is a random function. At time delay $\Delta t = 0$ it is identical to the mean fluctuation amplitude of the current, i.e.,

$$C(0) = \langle \Delta I^2(t) \rangle > 0.$$

The quantity $\sqrt{C(0)}$ is called the mean current noise. It is measured in [A].

For large time delays Δt the correlation function decays, i.e.,

$$\lim_{\Delta t \to \infty} C(\Delta t) = 0.$$

This expresses the fact that the value of the current noise at a certain instant is completely uncorrelated to a value far in the past (or in the future). It can be shown that, in general, the correlation function has the properties $|C(\Delta t)| \leq C(0)$, and $C(\Delta t) = C(-\Delta t)$. Often, the correlation function decays approximately exponentially according to

$$C(\Delta t) = C(0) e^{-|\Delta t|/\tau_c} = \left\langle \Delta I^2 \right\rangle e^{-|\Delta t|/\tau_c} \tag{20.2}$$

where τ is the characteristic decay constant, also called the correlation time. Such a decay is schematically shown in Fig. 20.2.

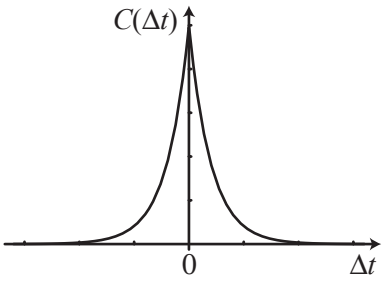


Fig. 20.2 Exponential decay of the autocorrelation function $C(\Delta t)$.

Spectral density. The Fourier transform of the correlation function is the spectral density, i.e.,

$$\tilde{S}(\omega) = \int_{-\infty}^{+\infty} dt C(t) e^{-i\omega t} = 2 \int_0^{+\infty} dt C(t) \cos(\omega t), \tag{20.3}$$

$$C(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega \tilde{S}(\omega) e^{i\omega t} = \frac{1}{\pi} \int_0^{+\infty} d\omega \tilde{S}(\omega) \cos(\omega t). \tag{20.4}$$

The spectral density is measured in units [A^2/Hz]. It is real valued and symmetric in ω. In apparatus measuring the spectral density of a signal, it is usually not $\tilde{S}(\omega)$, which is also defined for $\omega < 0$, that is displayed, but rather

$$S(\nu) = 2\tilde{S}(2\pi\nu), \tag{20.5}$$

where $\nu = \omega/2\pi$ is the frequency. The reason is that then, the mean noise current can be expressed as

$$\langle \Delta I \rangle^2 = C(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega \tilde{S}(\omega) = \int_0^{+\infty} d\nu S(\nu), \tag{20.6}$$

i.e., by the frequency integral of the measured spectral density. In the following, we will always use this measurement-related spectral density definition.

If the spectral density is a constant S_0 over a large frequency range, we talk about *white noise*. Thermal noise and shot noise can be classified as white noise, whereas $1/f$-noise is not white noise.

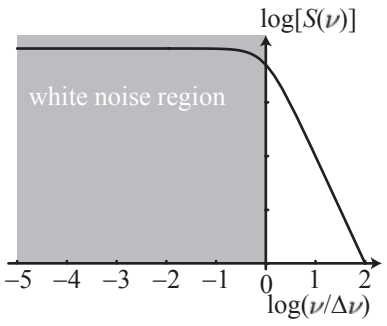


Fig. 20.3 Double logarithmic plot of the frequency-dependent noise spectral density for a system with an exponential autocorrelation function. The range of frequencies, where the noise spectral density is independent of frequency (white noise) is indicated in gray.

Spectral density of an exponentially decaying correlation function. If the correlation function $C(\Delta t)$ decays exponentially with $|\Delta t|$ [see eq. (20.2)], the spectral density is, according to eqs (20.3) and (20.5), given by

$$S(\nu) = 4\int_0^\infty dt C(0) e^{-t/\tau_c}\cos(2\pi\nu t) = \frac{4C(0)\tau_c}{1+(2\pi\nu\tau_c)^2} := \frac{S_0}{1+(2\pi\nu\tau_c)^2}. \quad (20.7)$$

This function is depicted in Fig. 20.3. It is apparent that the correlation time τ_c leads to first order low-pass cut-off of the spectral density at the bandwidth $\Delta\nu = 1/2\pi\tau_c$. For $\nu \ll \Delta\nu$ the noise is white with the constant spectral density $S_0 = 4C(0)\tau_c$.

Wiener–Khinchin relations. The spectral density can be related to the Fourier components of the noise current. To this end we transform, in the definition of the correlation function (20.1), the noise currents into the frequency domain

$$\Delta I(t) = \frac{1}{2\pi}\int d\omega i(\omega) e^{i\omega t},$$

and obtain for the correlation function

$$C(\Delta t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \left\langle |i(\omega)|^2 \right\rangle e^{i\omega\Delta t}.$$

Comparing with eq. (20.4) and using eq. (20.5) the spectral density of the current noise is therefore given by

$$S(\nu) = 2\langle |i(2\pi\nu)|^2\rangle, \quad (20.8)$$

i.e., by the averaged squared modulus of the Fourier spectrum of the noise current. Note that the average $\langle |i(\omega)|^2\rangle$ has units [A^2/Hz], whereas $|i(\omega)|^2$ has units [A^2s^2]. This is because the averaging procedure involves a frequency integration. It is given by

$$\left\langle |i(\omega)|^2 \right\rangle = \lim_{T\to\infty}\int \frac{d\omega'}{2\pi} i(\omega) i^\star(\omega') \frac{\sin[(\omega-\omega')T]}{(\omega-\omega')T}.$$

The factor $\sin x/x$ in the integrand becomes extremely sharp in the limit of large times and therefore essentially only $\omega' = \omega$ contributes to the frequency integral.

Probability density distribution of the noise amplitude. So far we have characterized the current noise in the time- and frequency-domain. Another function that is used to characterize noise is the probability density distribution of the noise current amplitude. Imagine that we measure the noise current at discrete points t_n (n integer) in time. We further split the current axis in discrete intervals of width δI. For a

given noise current $I(t)$ we can then determine a probability distribution for the noise current being in an interval δI around a specific value I. In the limit of very small intervals δI we obtain a probability density $p(I)dI$.

A typical probability density is the gaussian distribution with a certain width around a mean value. In this case we talk about gaussian noise. Vernon D. Landon realized in 1941 that, in electronics, noise signals originating from different sources cannot be distinguished by any known test procedure. Motivated by this observation he was able to show that the joint action of many independent small noise sources with arbitrary distribution functions will always result in a gaussian probability density distribution. This result is closely related to the central limit theorem. In this way nature produces, in many cases, gaussian noise.

The probability density distribution of the noise amplitude gives no information about the correlation function of the current. In the case of a constant spectral density (white noise) with a gaussian amplitude distribution we talk about gaussian white noise.

20.3 Filtering and bandwidth limitation

Usually, current (or other measured signals) are acquired with apparatus allowing only frequency components within a certain frequency interval to pass. The measurement device acts as a frequency filter and limits the bandwidth of the measurement. Bandwidth limitation usually leads to noise reduction. Below we will discuss how filtering and bandwidth limitation act on the measured signal and the measured noise power spectral density.

General considerations. A filter often acts on a time-dependent signal in real time. The filter is characterized by a time-dependent pulse response function $g(t)$ acting on the measured signal $I_{\text{in}}(t)$, producing an output signal $I_{\text{out}}(t)$ according to

$$I_{\text{out}}(t) = \int_{-\infty}^{\infty} dt' I_{\text{in}}(t') g(t - t').$$

An example would be the filter function

$$g(t) = \begin{cases} 1/t_0 & \text{for } 0 \leq t \leq t_0 \\ 0 & \text{elsewhere} \end{cases}$$

which describes data averaging over a fixed time span t_0. The filter function guarantees also that the current noise ΔI is transformed according to

$$\Delta I_{\text{out}}(t) = \int_{-\infty}^{\infty} dt' \Delta I_{\text{in}}(t') g(t - t').$$

In order to obtain the power spectral density of the filtered data, we Fourier transform this equation into the frequency domain and obtain

$$i_{\text{out}}(\omega) = i_{\text{in}}(\omega) g(\omega),$$

where $g(\omega)$ is the Fourier transform of the pulse response function $g(t)$. The power spectral density of the output signal is, according to eq. (20.8), given by

$$S_{\rm out}(\nu) = 2\langle |i_{\rm out}(2\pi\nu)|^2\rangle = 2\langle |i_{\rm in}(2\pi\nu)|^2\rangle |g(2\pi\nu)|^2 \\ = S_{\rm in}(\nu)|g(2\pi\nu)|^2. \quad (20.9)$$

The action of the filter on the measured spectral density is therefore the product of the input spectral density and the squared magnitude of the filter response function in the frequency domain.

Filter bandwidth. The bandwidth $\Delta\nu_{\rm BW}$ of a filter is usually defined as the width of the filter function $|g(2\pi\nu)|^2$ at half the maximum value (between the -3dB points). For filters that have their maximum value at zero frequency the bandwidth is the frequency at which the response function has decayed to half of the maximum value (-3 dB point).

Example I: averaging over finite time span. If we continue the discussion of the above time averaging filter function we find the Fourier transform

$$|g(\omega)|^2 = \frac{\sin^2(\omega t_0/2)}{(\omega t_0/2)^2}.$$

This filter function decays with increasing frequency on the scale $2/t_0$. The oscillating numerator has zeros at frequencies $\nu = n/t_0$ (n integer). It therefore has the property to block these frequencies completely. If we assume that $S_{\rm in}(\nu) = S_0$ for frequencies ν below and much beyond $1/t_0$ (white noise), then the output noise is, according to (20.6), given by

$$\langle \Delta I\rangle^2 = \frac{S_0}{\pi t_0}\int_0^\infty dx \frac{\sin^2 x}{x^2} = \frac{S_0}{2t_0} := S_0\Delta\nu,$$

where we call

$$\Delta\nu = \frac{1}{2t_0} \quad (20.10)$$

the equivalent noise bandwidth of the filter. It describes the width of a fictitious rectangular filter in the frequency domain such that the noise power in this rectangular band is equal to the actual output power. The above result has the well-known implication that reducing the measurement bandwidth by increasing t_0 reduces the output noise. According to our above definition, this filter has a bandwidth of $\Delta\nu_{\rm BW} = 1.39/\pi t_0 = 0.88\Delta\nu$.

Example II: First-order low-pass filter. If white noise is measured with apparatus having a first-order low-pass characteristic, the pulse response function in real time is given by

$$g(t) = \begin{cases} \tau^{-1}e^{-t/\tau} & \text{for } t \geq 0 \\ 0 & \text{elsewhere} \end{cases},$$

where τ is the correlation time. For the Fourier transform we find

$$|g(\omega)|^2 = \frac{1}{1+(\omega\tau)^2},$$

and therefore $\Delta\nu_{\mathrm{BW}} = 1/2\pi\tau$ is the bandwidth of the filter. At high frequencies, the filter cuts off proportional to ω^{-2} corresponding to a damping of 6 dB/octave. If we assume that we filter a white noise spectrum with spectral density $S(\nu) = S_0$, the measured current noise is, according to (20.6), given by

$$\langle \Delta I \rangle^2 = \int_0^{+\infty} d\nu \frac{S_0}{1+(\nu/\Delta\nu_{\mathrm{BW}})^2} = S_0 \frac{\pi\Delta\nu_{\mathrm{BW}}}{2} := S_0\Delta\nu. \qquad (20.11)$$

The equivalent noise bandwidth is here $\Delta\nu = \Delta\nu_{\mathrm{BW}}\pi/2$.

Example III: first order low-pass acting on noise with exponential correlation function. As a further example we calculate the noise in a measurement where the ammeter shows a first-order low-pass characteristic with bandwidth $\Delta\nu_{\mathrm{BW}}$, and the correlation function of the input spectral density has an exponential decay with correlation time τ_{c} (equivalent to a low-pass bandwidth $\Delta\nu_{\mathrm{c}} = 1/2\pi\tau_{\mathrm{c}}$) leading to the spectral density $S(\nu)$ of eq. (20.7). The measured noise will then be

$$\left\langle \Delta I^2 \right\rangle = \int_0^{\infty} d\nu S(\nu) \frac{1}{1+(\nu/\Delta\nu)^2} = \frac{S_0}{1+\Delta\nu_{\mathrm{BW}}/\Delta\nu_{\mathrm{c}}} \frac{\pi\Delta\nu_{\mathrm{BW}}}{2}. \quad (20.12)$$

In the case of $\Delta\nu_{\mathrm{BW}} \ll \Delta\nu_{\mathrm{c}}$ we recover the result of eq. (20.11)

$$\left\langle \Delta I^2 \right\rangle_{\Delta\nu} = S_0\Delta\nu \qquad (20.13)$$

with the equivalent noise bandwidth $\Delta\nu = \pi\Delta\nu_{\mathrm{BW}}/2$.

It is also of practical interest to regard the input spectral density $S(\nu)$ as the output of a first stage low-pass filter with bandwidth $\Delta\nu_{\mathrm{c}}$. Together with the additional low-pass filter this corresponds to a two-stage low-pass filter. We obtain a so-called second-order low-pass filter in the case where both stages have the same bandwidth, e.g., $\Delta\nu_{\mathrm{c}} = \Delta\nu_{\mathrm{BW}}$. The equivalent noise bandwidth of this second-order low-pass can be read to be $\Delta\nu = \pi\Delta\nu_{\mathrm{BW}}/4$. In fact, it is easy to verify that using even more low-pass filtering stages reduces the equivalent noise bandwidth even more. Equivalent noise bandwidths for low-pass filters of certain orders n are tabulated in Table 20.1 (final column).

Step response time and equivalent noise bandwidth of higher order low-pass filters. The price to pay for the noise reduction with low-pass filters of increasing order is an increase of the effective correlation time, i.e., a slower response in the time-domain. This can, for example, be seen in the step response. Assume the input current is characterized by a unit step at time zero. How long does it take for the

Table 20.1 Delay times of the step response, and equivalent noise bandwidths for low-pass filters of order n, time constant τ and bandwidth $\Delta\nu = 1/2\pi\tau$. The second to third column gives delay times to reach the indicated percentage of the full step height in units of τ. The final column is the equivalent noise bandwidth in units of $\Delta\nu_{\mathrm{BW}}$ of the first order low-pass.

n	50%	90%	98%	$\Delta\nu/\Delta\nu_{\mathrm{BW}}$
1	0.69	2.30	3.91	$\pi/2$
2	1.68	3.89	5.83	$\pi/4$
3	2.67	5.32	7.52	$3\pi/16$
4	3.67	6.68	9.08	$5\pi/32$

output signal to reach the value of one? For a first-order low-pass filter we find the step response

$$I_{\text{out}}(t) = 1 - e^{-t/\tau}.$$

This means that the output value of one is never reached and our original question does not make much sense. However, we can ask how long it takes for the output signal to reach a certain percentage of the full step height, such as 50%, or 90%. The corresponding delay times in units of τ for three percentages are given in Table 20.1.

20.4 Thermal noise

After having discussed quantities characterizing noise, and the action of noise filtering techniques, we turn our attention to fundamental physical origins of noise. We start with a discussion of thermal noise.

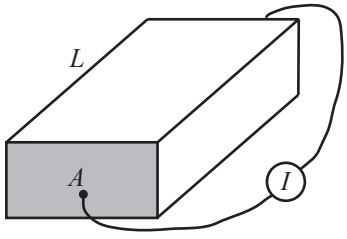


Fig. 20.4 Piece of material with cross-sectional area A and length L acting as a resistor.

Johnson–Nyquist noise of a resistor. We consider a resistor with resistance R where both sides are shorted, for example, via an (ideal) ammeter, as depicted in Fig 20.4. For the sake of simplicity we assume it to consist of a rectangular piece of material with cross-sectional area A and length L. We assume the system to be in thermodynamic equilibrium with its environment, characterized by a constant temperature T. In the statistical average, the same amount of electrons will move in the x- and in the $-x$-direction, and the net current is exactly zero. During short time intervals, however, it can accidentally happen that a few more electrons move in one direction than in the other. This leads to a short-term current in one direction which averages to zero with short-term currents in the other direction over long times. These short-term statistical fluctuations of the current originate from the thermal motion of the electrons.

In order to describe this setting mathematically, we first calculate the current contribution in the $+x$-direction of an individual electron. From the current density $j_x = -|e|nv_x$ for a system with electron density n we deduce, for a single electron, a current contribution

$$I_1(t) = \frac{-|e|v_x(t)}{L}.$$

Since the average current $\langle I_1(t)\rangle$ for a shorted resistor is zero, the correlation function of the current of a single electron is given by

$$C_1(\Delta t) = \langle I_1(t)I_1(t+\Delta t)\rangle = \frac{e^2}{L^2}\,\langle v_x(t)v_x(t+\Delta t)\rangle\,.$$

The velocity correlation function in a resistor characterized by a Drude scattering time τ is given by $\langle v_x(t)v_x(t+\Delta t)\rangle = \langle v_x^2(0)\rangle \exp(-\Delta t/\tau)$. Considering the motion of all N electrons in the resistor as being inde-

pendent we therefore obtain the correlation function

$$C(\Delta t) = NC_1(\Delta t) = \frac{Ne^2}{L^2}\left\langle v_x^2(0)\right\rangle e^{-\Delta t/\tau}$$
$$= \frac{ne^2\tau}{m^\star}\frac{A}{L}\frac{m^\star\left\langle v_x^2(0)\right\rangle}{\tau}e^{-\Delta t/\tau} = G\frac{m^\star\left\langle v_x^2\right\rangle}{\tau}e^{-\Delta t/\tau},$$

where we have used the Drude expression for the conductivity, and $G = 1/R$ is the conductance.

The crucial step is now to incorporate the relation between the average kinetic energy of an electron and the temperature of the system. According to the equipartition theorem of thermodynamics, each classical degree of freedom possesses a mean energy of $k_\mathrm{B}T/2$ in thermodynamic equilibrium. For our case this means that $m^\star\left\langle v_x^2\right\rangle/2 = k_\mathrm{B}T/2$. As a result we find

$$C(\Delta t) = Gk_\mathrm{B}T\frac{1}{\tau}e^{-\Delta t/\tau}.$$

This is an exponential correlation function as introduced in eq. (20.2), where the scattering time τ plays the role of the correlation time. We can therefore use eq. (20.7) to find the thermal noise spectral density

$$S(\nu) = \frac{4Gk_\mathrm{B}T}{1+(2\pi\nu\tau)^2}. \tag{20.14}$$

For frequencies $\nu \ll 1/2\pi\tau$ the spectral density describes white noise with $S_0 = 4Gk_\mathrm{B}T$ and the noise within a frequency band $\Delta\nu$ is

$$S_0\Delta\nu = 4Gk_\mathrm{B}T\Delta\nu.$$

This equation is known as the Johnson–Nyquist formula for thermal noise.

The existence of thermal noise implies that every physical measurement will suffer from noise. Each sample with a finite electrical resistance will produce the corresponding thermal noise. In Table 20.2 we list the spectral density S_0 for a series of different resistors at room temperature.

Table 20.2 Spectral density $S_0 = 4Gk_\mathrm{B}T$ of the current noise of ohmic resistors at room temperature.

$R\ (\Omega)$	$S_0\ (\mathrm{A}^2/\mathrm{Hz})$
1	1.63×10^{-20}
10^3	1.63×10^{-23}
10^6	1.63×10^{-26}

Relation with the fluctuation–dissipation theorem. The thermal noise is a direct consequence of the relation between the fluctuations of thermodynamic quantities (here: the current), and the linear response functions (here: the conductivity). This relation is known in thermodynamics as the fluctuation–dissipation theorem.

What information about the resistor can we obtain from the measurement of the thermal noise of its resistance? If the temperature of the environment is known, the measurement of thermal noise does not give information beyond the value of the resistor itself, and is therefore less interesting than shot noise, as we will see below. However, if the resistance is known (and it is easy to measure it), the thermal noise can be used for operating a resistor as a primary thermometer, or for measuring the electronic temperature (which may differ, for example at liquid He temperatures, from the temperature of the crystal lattice).

20.5 Shot noise

While the thermal noise does not give access to any other system properties than the conductance, this is not true for measurements of the shot noise. The latter is sometimes also called *excess noise*. It contains information about the temporal correlations of the electrons which are not contained in the average current and therefore not in the conductance. In general, shot noise may contain

- information about the charge of particles that contribute to the current,
- information about the statistics obeyed by the particles contributing to the current (Fermi–Dirac statistics or Bose–Einstein statistics),
- information about interactions causing correlations between particles, and
- information about the transmission $\mathcal{T}$ (which can also be obtained from conductance measurements).

20.5.1 Shot noise of a vacuum tube

As a first example for a system in which shot noise arises we consider a vacuum tube [see Fig. 20.5(a)], in which thermally excited electrons are emitted from the hot cathode, and sucked away by a large electric field. The tunneling barrier is characterized by a transmission $\mathcal{T}(E)$ depending on the energy of the impinging electron [see Fig. 20.5(b)]. Well below the top of the barrier the transmission is exponentially suppressed, whereas far above the barrier it is essentially one. Close to the top of the barrier the transmission exhibits a sharp step from values $\mathcal{T}(E) \ll 1$ to $\mathcal{T}(E) \approx 1$. The work function W of the cathode material is of the order of 5 eV, the cathode temperature is of the order of 2000°C corresponding to a thermal energy $k_{\mathrm{B}}T \approx 190\,\mathrm{meV}$. The occupation $f_{\mathrm{K}}(E)$ of states in the cathode is given by the Fermi–Dirac distribution function. Since $k_{\mathrm{B}}T \ll W$ the occupation probability close to the barrier top, where the

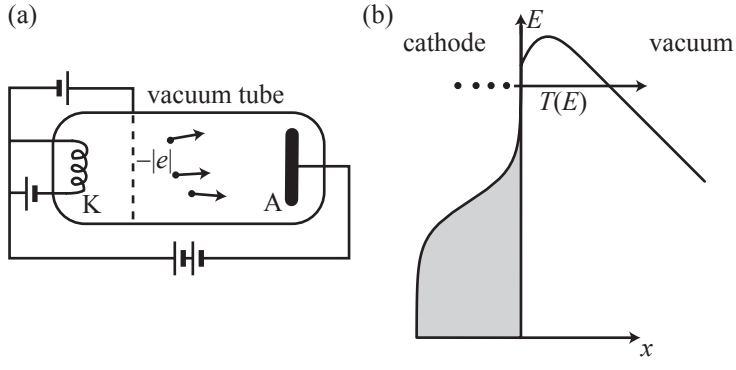


Fig. 20.5 (a) Schematic diagram of a vacuum tube with a hot cathode (K) and an anode (A). (b) Schematic diagram thermally activated tunneling from the anode of the diode into the vacuum.

transmission becomes appreciable, is very small and well described by the Boltzmann distribution. The states on the vacuum side of the barrier can be considered to be unoccupied because any tunneling electron is immediately sucked away. Thermionic emission of electrons from the cathode is determined by the interplay between the sharp step of the transmission function at the barrier top, and the exponential tail of the Boltzmann distribution. As a result, the product $f_{\mathrm{K}}(E)\mathcal{T}(E)$ shows a marked maximum near the top of the barrier, but even there the value $f_{\mathrm{K}}(W)\mathcal{T}(W) \ll 1$ (we choose the cathode Fermi energy as the energy zero). This is the characteristic situation for thermionic emission of electrons. The number of cathode states at this energy is proportional to the density of states $\mathcal{D}(W)$.

Electron transmission as a probabilistic experiment. We now describe the electron emission process from the cathode as a probabilistic experiment. Assume that within an observation time t_0, $N \propto \mathcal{D}(W)$ attempts were possible for electrons to hit the barrier. The statistics of whether such a potential attempt leads to a tunneling electron or not depends on the probability $p = f_{\mathrm{K}}(W)\mathcal{T}(W)$. The situation for an individual potential attempt is the same as in a probabilistic experiment with two possible outcomes, such as tossing a coin. Here, the two outcomes are, (1) an attempt is successful (probability p) and an electron is transmitted, or (2) an attempt is unsuccessful (probability $1-p$) and no electron is transmitted. The probability that out of the N attempts, n electrons are transmitted, is then given by the *binomial distribution*

$$P(n) = \left(\begin{array}{c} N \\ n \end{array} \right) p^n (1-p)^{1-n} = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}, \qquad (20.15)$$

which is shown in Fig. 20.6 for the particular case of $N = 30$ and $p = 0.3$. The distribution has the property that the expectation value $\langle n \rangle = Np$, and the variance of the distribution is $\sigma^2 = \langle (n - \langle n \rangle)^2 \rangle = Np(1-p)$.

In the limit of $p \ll 1$ the binomial distribution can be well approximated by the *Poisson distribution*

$$P(n) = \frac{\mu^n}{n!} e^{-\mu}, \qquad (20.16)$$

with the mean value $\mu \equiv \langle n \rangle = Np$ and the variance $\sigma^2 = \mu = Np$. An example of the Poisson distribution function is shown in Fig. 20.7 for parameters $N = 30$ and $p = 0.05$, where it is already a reasonable approximation for the binomial distribution. Because in our example the probability $p \ll 1$ at relevant energies around W, the Poisson distribution in eq. (20.16) can be seen as the counting statistics of the transmitted electrons.

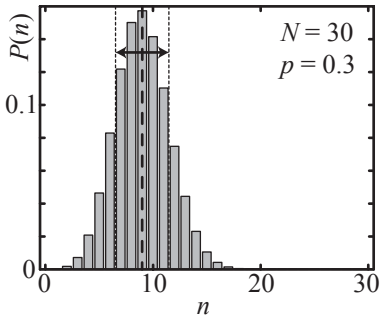


Fig. 20.6 Binomial distribution for $N = 30$ and $p = 0.3$. The vertical thick dashed line indicates the position of the expectation value $\langle n \rangle = Np = 9$. The two thinner vertical dotted lines indicate the width $2\sigma \approx 5$ of the distribution.

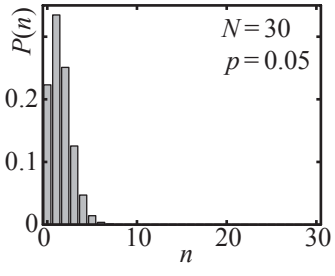


Fig. 20.7 Poisson distribution function for parameters $N = 30$ and $p = 0.05$.

Average current and classical shot noise. With the Poisson distribution function, the mean electrical current is calculated to be

$$I = -\frac{|e|\,\langle n \rangle}{t_0} = -\frac{|e|N}{t_0} p.$$

It could be measured in a way in which the transmitted electrons are repeatedly counted over a time span t_0 on the anode side. The average of the number of counts is then determined from the measured counting statistics which we have assumed to be poissonian. The time span t_0 plays the role of an integration time, or equivalently, $\Delta\nu = 1/2t_0$ is the bandwidth of the measurement apparatus [cf., eq. (20.10)].

The shot noise is revealed if we consider the temporal fluctuations in the number of transmitted electrons which is related to the width of the distribution function in Fig. 20.7. According to the results of probability theory, the average of these fluctuations is given by

$$(\Delta n)^2 = \left\langle (n - \langle n \rangle)^2 \right\rangle = \mu = Np.$$

Correspondingly, the mean fluctuations of the electrical current are

$$\left\langle \Delta I^2 \right\rangle_{t_0} = \left\langle I^2 \right\rangle - \langle I \rangle^2 = \frac{e^2 (\Delta n)^2}{t_0^2} = \frac{e^2}{t_0^2} Np = \frac{|e|}{t_0} \left| \langle I \rangle \right| = 2|e| \left| \langle I \rangle \right| \Delta\nu.$$

We therefore find the spectral density of the shot noise

$$S_0 = 2|e| \left| \langle I \rangle \right|. \tag{20.17}$$

This relation is known as the classical shot noise formula, or the *Schottky formula*. It expresses the fact that current flow in the vacuum tube under the conditions outlined above is inevitably causing current noise which increases proportional to the magnitude of the current.

20.5.2 Landauer's wave packet approach

We have seen that shot noise is a nonequilibrium phenomenon which arises because charge is transported in discrete elementary portions, usually of the elementary charge $|e|$, through a conductor. The effect is similar to the noise of rain drops impinging randomly on a plane surface. Shot noise emphasizes the particle character of the quantum mechanical charge carriers. The quantum description that comes closest to the particle character uses wave packets that can be constructed, for example, from plain wave states. In order to demonstrate such a construction of wave packets we choose a one-dimensional example.

Decomposition of plane wave states into a train of wave packets in one dimension. Following the approach of Martin and Landauer, 1992, we decompose the current contribution of incident plane wave states in an energy interval ΔE of a one-dimensional channel into wave packets. The energy interval ΔE is chosen to be reasonably small such that, for example, the transmission of the structure or the involved Fermi distribution functions do not change appreciably over this energy interval. The wave packets moving in the positive x-direction are constructed from plane waves

$$\psi(x,t) = \sqrt{\frac{m^\star}{\hbar k}} e^{i(kx - Et/\hbar)}.$$

We assume a parabolic dispersion relation $E(k) = \hbar^2 k^2/2m^\star$, and $k = \sqrt{2m^\star E/\hbar^2}$. The prefactor ensures that each wave contributes the same unit particle flux density in the x-direction. The wave packet n is formed by a superposition of states in the energy interval $[E - \Delta E, E + \Delta E]$ according to

$$\psi_n(x,t) = \frac{1}{\sqrt{2h\Delta E}} \int_{E-\Delta E}^{E+\Delta E} dE' \sqrt{\frac{m^\star}{\hbar k(E')}} e^{ik(E')x - iE'(t+n\tau)/\hbar},$$

where $\tau = h/2\Delta E$, and n is an integer. A single wave packet is depicted in Fig. 20.8. It moves with the group velocity $v_g = \hbar k/m^\star$. All these wave packets form an orthonormal basis of states for this energy interval. Orthonormality is seen by calculating

$$\int dx \psi_n^\star(x,t)\psi_m(x,t) = \frac{1}{2h\Delta E} \int_{E-\Delta E}^{E+\Delta E} dE' \sqrt{\frac{m^\star}{\hbar k(E')}} \int_{E-\Delta E}^{E+\Delta E} dE'' \sqrt{\frac{m^\star}{\hbar k(E'')}} \times \int dx e^{i(k(E'')-k(E'))x - i(E''-E')t/\hbar - i(E''m - E'n)\tau/\hbar}.$$

The integration over x can be performed and leads to $2\pi\delta(k(E') - k(E''))$. The integration over E'' can be performed after transformation into an integration over k''. This eventually leads to

$$\int dx \psi_n^\star(x,t)\psi_m(x,t) = \frac{1}{2\Delta E} \int_{E-\Delta E}^{E+\Delta E} dE' e^{-iE'(m-n)\tau/\hbar}.$$

It is evident from this result that for $m = n$

$$\int dx \psi_n^\star(x,t)\psi_n(x,t) = \frac{1}{2\Delta E} \int_{E-\Delta E}^{E+\Delta E} dE' = 1,$$

and for $m \neq n$

$$\int dx \psi_n^\star(x,t)\psi_m(x,t) = \frac{e^{-iE(m-n)\tau/\hbar}}{\Delta E(m-n)\tau/\hbar} \sin\left[\Delta E(m-n)\tau/\hbar\right].$$

Inserting the expression for τ we obtain for the case $m \neq n$

$$\int dx \psi_n^\star(x,t)\psi_m(x,t) = \frac{e^{-i\pi E(m-n)/\Delta E}}{2\pi(m-n)} \sin\left[\pi(m-n)\right] = 0,$$

because $m - n$ is an integer number.

This example shows that we can describe the set of states in the energy interval $[E - \Delta E, E + \Delta E]$ traveling in the positive x-direction not only using a set of plane wave states, but also using a pulse train made of wave packets traveling in this direction. Wave packets in the train have equidistant spacing in time (τ) and space ($v_g\tau$). The series of wave packets with all integer numbers n forms a pulse train as depicted in Fig. 20.9. Neighboring wave packets will overlap, but they are orthogonal. Owing to the Pauli principle each packet can be occupied only with two electrons (spin up and down).

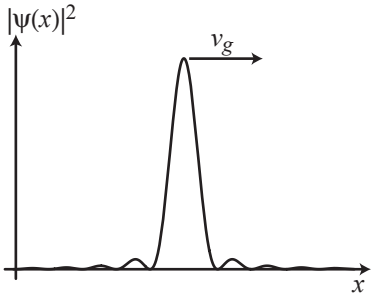


Fig. 20.8 Wave packet composed of plane wave states within an energy interval $2\Delta E$.

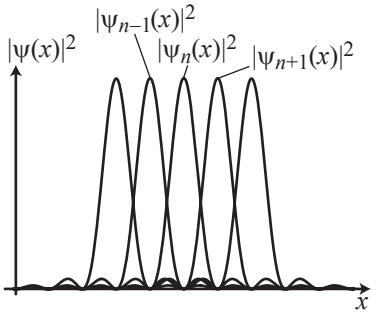


Fig. 20.9 Pulse train of wave packets each carrying a maximum of two electrons.

From wave packets to current pulses. Following Martin and Landauer, 1992, we now make the conceptual step from wave packets (wave functions) to current pulses. Although this step neglects interference terms arising from the fermionic statistics of the electrons, and may lead to problems in some cases, it illustrates the basic ideas behind shot noise in a very intuitive way. In this picture, wave packet n corresponds to a current pulse $j(t - n\tau)$. The total current of a stream of electrons at a specific energy E can be written as

$$I(t) = \sum_n j(t - n\tau) g_n,$$

where $g_n = 0, 1$, depending on whether the wave packet is occupied or not. This is the basic idea behind Landauer's wave packet approach.

20.5.3 Noise of a partially occupied monoenergetic stream of fermions

An example where the noise is governed by binomial statistics is that of a single one-dimensional channel of electrons propagating at energies around E in one direction. In order to calculate the noise in such a stream, we write the noise component of the current as

$$\Delta I(t) = \sum_n j(t - n\tau)(g_n - \langle g_n \rangle),$$

where $g_n = 0, 1$ indicating an occupied, or an empty current pulse (wave packet). We assume that the average $\langle g_n \rangle$, i.e., the average fraction of occupied wave packets, is equal to a given probability p which is between zero and one. We now calculate the spectral noise density using eq. (20.8). To this end, we determine the Fourier transform of the current fluctuations to be

$$i(\omega) = \sum_n (g_n - \langle g_n \rangle) \int dt\; j(t - n\tau) e^{-i\omega t} = j(\omega) \sum_n (g_n - \langle g_n \rangle) e^{-i\omega n\tau}.$$

The spectral density is then found to be

$$\begin{aligned} S(\nu) &= 2\langle |i(2\pi\nu)|^2 \rangle \\ &= 2|j(2\pi\nu)|^2 \left\langle \sum_{n,n'} (g_n - \langle g_n \rangle)(g_{n'} - \langle g_n \rangle) e^{-i2\pi\nu(n-n')\tau} \right\rangle \\ &= 2\frac{|j(2\pi\nu)|^2}{\tau} \langle (g_n - \langle g_n \rangle)^2 \rangle. \end{aligned}$$

We are interested here in low frequencies for which $2\pi\nu\tau \ll 1$. We can therefore replace $j(2\pi\nu) \to j(0) = -|e|$. The required average can be worked out as follows:

$$\langle (g_n - \langle g_n \rangle)^2 \rangle = \langle g_n^2 \rangle - \langle g_n \rangle^2 = 1^2 \cdot p + 0^2 \cdot (1 - p) - p^2 = p(1 - p).$$

This factor is the variance of the binomial distribution (20.15). Inserting these result gives the spectral density

$$S(\nu) = \frac{2e^2}{h} p(1-p)\Delta E = 2|e||\Delta I(E)|(1-p), \tag{20.18}$$

where $\Delta I = e^2 p \Delta E/h$ is the current associated with the stream of fermions. This result tells us that the shot noise of such a stream of fermions is symmetric around the value $p = 1/2$, where the noise is maximum. The shot noise decays to zero as $p \to 0$, or $p \to 1$. Such a stream of fermions with $0 < p < 1$ is called partitioned.

20.5.4 Zero temperature shot noise with binomial distribution

A practical application of the above result arises if tunneling of electrons in a nanostructure at very low temperature is considered. An example would be the quantum point contact with a conductance anywhere below the first conductance plateau (lowest mode starts to transmit). If a voltage $V_{\rm SD}$ is applied between source and drain of a quantum point contact structure which is much larger than temperature $k_{\rm B}T$, then electrons can tunnel through the barrier only within the energy window (bias window) $|e|V_{\rm SD}$, as a direct consequence of the Pauli principle. Temperature smearing can be neglected. Using the uncertainty relation we can estimate the rate of electrons hitting the barrier to be $\tau^{-1} = eV_{\rm SD}/h$. Within a time interval t_0 there are therefore $N = eVt_0/h$ attempts to tunnel. At zero temperature there are no fluctuations of this number, because each wave packet is, according to the Pauli principle, occupied with exactly one electron. The probability that a particular attempt to tunnel is successful is given by the transmission probability $\mathcal{T}$ which can take arbitrary values between zero and one. Behind the tunneling barrier the stream of electrons is partitioned with $p = \mathcal{T}$. According to eq. (20.18) the spectral density of the noise is then given by

$$S(\nu) = \frac{2e^2}{h} \mathcal{T}(1-\mathcal{T})|e|V_{\rm SD} = 2|e||\langle I\rangle|(1-\mathcal{T}). \tag{20.19}$$

It differs by the factor $(1-\mathcal{T})$ from the classical result of Poisson statistics. However, in the limit $\mathcal{T} \ll 1$, both results are identical, because $1 - \mathcal{T} \approx 1$.

The ratio of the spectral densities $S(\nu)/S_{\rm Schottky}(\nu)$ is called the *Fano factor*:

$$F = \frac{S(\nu)}{S_{\rm Schottky}(\nu)}.$$

In our example, $F = 1 - \mathcal{T}$ is smaller than one, meaning that the quantum mechanical shot noise at large transmission probabilities $\mathcal{T}$ is suppressed compared to the classical shot noise. If the transmission $\mathcal{T} \to 1$, the quantum shot noise even goes to zero, like the Fano factor.

20.6 General expression for the noise in mesoscopic systems

In order to derive a general expression for the noise in noninteracting mesoscopic systems with an arbitrary number of transmission channels, and at arbitrary source–drain voltages and temperatures, we assume that

- the noise of individual transmission channels is additive, i.e., their noise is statistically independent,
- the noise of individual energy intervals is additive (i.e., statistically independent), and
- the two spin orientations fluctuate independently from each other.

We imagine a pulse train of wave packets of energy E to impinge from the source contact on the mesoscopic system under consideration. The fraction of packets that are occupied in the incoming pulse train is given by the Fermi–Dirac distribution function $f_{\mathrm{S}}(E)$ in the source. At the same time, there will be a pulse train of wave packets at the drain side moving towards the mesoscopic system under consideration. For these states, the fraction of occupied packets is given by $f_{\mathrm{D}}(E)$. As pointed out by Martin and Landauer, 1992, it is convenient to construct the wave packets on both sides in such a way that they are synchronized. This makes sure that a transmitted wave packet from the source maps into the same state as a reflected wave packet from the drain.

Each occupied wave packet propagating in the structure constitutes a current pulse $j(t - n\tau)$. If we were to measure the current by counting the current pulses traversing an imaginary plane in, say, the drain contact, we would find a pulse pattern similar to that depicted in Fig. 20.10. Positive pulses originate from wave packets moving in the source–drain direction, negative pulses to those moving in the drain–source direction. The total time-dependent current at the specific energy E could be written as

$$\Delta I(t) = \sum_n j(t - n\tau) g_n. \tag{20.20}$$

Following the analysis of Martin and Landauer, 1992, there are six possible pulse histories:

(1) Two corresponding occupied wave packets are incident from the

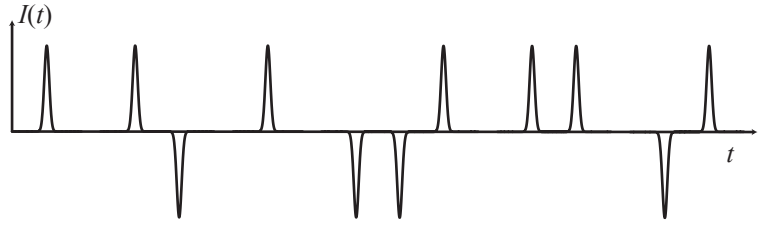


Fig. 20.10 Time-dependent current in a mesoscopic structure. The current pulses are individual electrons that are transmitted. A positive pulse means that an electron has moved from source to drain, a negative pulse indicates the opposite direction of motion.

source and the drain at the same time. The probability for this to occur is $f_S f_D$. In this case, no net current results, i.e., $g_n = 0$.

(2) Two corresponding unoccupied wave packets are incident from the source and the drain at the same time. The probability for this to occur is $(1 - f_S)(1 - f_D)$. Also in this case, no net current results ($g_n = 0$).

(3) Two corresponding occupied wave packets are incident from the source and the drain at the same time with the one on the source side occupied and that on the drain side empty. In this case, the electron can be transmitted. The probability for this to occur is $p_{DS} = f_S T(1 - f_D)$. In this case, a positive current pulse is measured ($g_n = 1$).

(4) Two corresponding occupied wave packets are incident from the source and the drain at the same time with the one on the source side occupied and that on the drain side empty. However, the electron is reflected. The probability for this to occur is $f_S(1 - T)(1 - f_D)$. In this case, the current is zero ($g_n = 0$).

(5) Two corresponding occupied wave packets are incident from the source and the drain at the same time with the one on the drain side occupied and that on the source side empty. In this case, the electron can be transmitted as well. The probability for this to occur is $p_{SD} = (1 - f_S)T f_D$. In this case, a negative current pulse is measured ($g_n = -1$).

(6) Two corresponding occupied wave packets are incident from the source and the drain at the same time with the one on the drain side occupied and that on the source side empty. However, the electron is reflected. The probability for this to occur is $f_D(1 - T)(1 - f_S)$. In this case, the current is zero ($g_n = 0$).

Average current. Equation (20.20) together with the above analysis allows us to calculate the average current at energy E through the mesoscopic device. We find

$$\langle \Delta I \rangle = -\frac{|e|\langle g_n \rangle}{\tau} = -\frac{|e|(p_{DS} - p_{SD})}{\tau} = -\frac{|e|}{h}\mathcal{T}(E)[f_S(E) - f_D(E)]\Delta E$$

in full agreement with the Landauer–Büttiker theory of conductance. Integrating over the energy gives the familiar expression for the total current

$$\langle I \rangle = -\frac{|e|}{h}\int dE\, \mathcal{T}(E)[f_S(E) - f_D(E)].$$

Current noise. We continue with the calculation of the current fluctuations at energy E given by

$$\delta I(t) = \sum_n j(t - n\tau) g_n + \frac{|e|}{\tau}\langle g_n \rangle .$$

The spectral density $dS(E;\omega)$ at the energy E is obtained from the mean square of the Fourier transformed current fluctuations at this energy (cf. 20.8) and it is found that

$$dS(E;\omega) = \frac{2e^2}{\tau}\left(\langle g_n^2\rangle - \langle g_n\rangle^2\right).$$

The averages of the g_n are calculated as above. After integration over the energy, and summation over spin and transmission channels, we obtain the total spectral density

$$\tilde{S}(\omega) = \frac{2e^2}{h}\sum_n \int_0^\infty dE\left\{[f_S(1-f_D)+f_D(1-f_S)]\,\mathcal{T}_n - [f_S-f_D]^2\,\mathcal{T}_n^2\right\}. \tag{20.21}$$

This expression contains the contribution of the thermal noise as well as that of the shot noise. There are a number of equivalent forms of this expression that occur in the literature, of which we quote only one, which is particularly nice to interpret:

$$\tilde{S}(\omega) = \frac{2e^2}{h}\sum_n \int_0^\infty dE\left\{[f_S-f_D]^2\mathcal{T}_n(1-\mathcal{T}_n) - k_BT[df_S/dE + df_D/dE]\mathcal{T}_n\right\}.$$

It separates the noise of two sources: one is proportional to k_BT and goes towards the thermal equilibrium noise if no source–drain voltage is applied. The other term vanishes in thermodynamic equilibrium, but it survives at $k_BT = 0$, if the source–drain voltage is nonzero. This is the shot noise term.

Thermal noise. We first discuss the spectral density for the case of zero source–drain voltage, i.e., $f_S = f_D \equiv f$. This is the case of thermodynamic equilibrium. The spectral density becomes

$$\tilde{S}(\omega) = \frac{4e^2}{h}\sum_n \int_0^\infty dE f(1-f)\mathcal{T}_n.$$

Using the property of the Fermi–Dirac distribution $f(1-f) = -k_BTdf/dE$ and the Landauer–Büttiker expression for the conductance, we obtain

$$\tilde{S}(\omega) = \frac{4k_BTe^2}{h}\sum_n \int_0^\infty dE\mathcal{T}_n(E)\left(-\frac{df(E)}{dE}\right) = 4k_BTG.$$

This expression is identical to the thermal noise formula (20.14), except for the finite bandwidth cut-off.

Shot noise. In order to extract the shot noise contribution to the spectral density, we consider the case of strong forward bias, as it is applied to a vacuum diode. In this case $f_D = 0$ and we obtain from eq. (20.21)

$$\tilde{S}(\omega) = \frac{2e^2}{h}\sum_n \int_0^\infty dE\left\{f_S\mathcal{T}_n(1-\mathcal{T}_n) + f_S(1-f_S)\mathcal{T}_n^2\right\}.$$

The distribution function f_S will be exponentially small at relevant energies, such that $1 - f_S \approx 1$. This leads to

$$\tilde{S}(\omega) = \frac{2e^2}{h} \sum_n \int_0^\infty dE \left\{ f_S \mathcal{T}_n (1 - \mathcal{T}_n) + f_S \mathcal{T}_n^2 \right\}.$$

In the case of *classical transmission* over the top of the barrier (work function W), we have $T = 1$ for electrons above the barrier and $T = 0$ for electrons below. As a result we obtain the classical Schottky formula

$$\tilde{S}(\omega) = \frac{2e^2}{h} \sum_n \int_W^\infty dE f_S = 2e|I|.$$

In the *quantum case* with $\mathcal{T}_n \ll 1$ we also find the Schottky formula

$$\tilde{S}(\omega) = \frac{2e^2}{h} \sum_n \int_0^\infty dE f_S(E) \mathcal{T}_n(E) = 2e|I|.$$

Beyond these very special conditions of the vacuum tube, we consider the shot noise at zero temperature. We then have

$$\tilde{S}(\omega) = \frac{2e^2}{h} \sum_n \int_\mu^{\mu + eV_{SD}} dE \mathcal{T}_n(E)[1 - \mathcal{T}_n(E)].$$

If we neglect the energy dependence of the transmission we obtain the well-known mesoscopic shot noise formula

$$\tilde{S}(\omega) = 2eV_{SD} \frac{e^2}{h} \sum_n \mathcal{T}_n(E_F)[1 - \mathcal{T}_n(E_F).$$

This expression simplifies in the limit of small transmission $\mathcal{T}_n \ll 1$ to the Schottky formula. It is remarkable that a transmission channel with $\mathcal{T}_n = 1$ does not contribute to the shot noise. The same is also valid for the case of $\mathcal{T}_n = 0$. For arbitrary values of the transmission between zero and one, the shot noise is suppressed compared to the Schottky formula by the factors $(1 - \mathcal{T}_n)$. The Fano factor is in this general case given by

$$F = \frac{\sum_n \mathcal{T}_n (1 - \mathcal{T}_n)}{\sum_n \mathcal{T}_n}. \tag{20.22}$$

20.7 Experiments on shot noise in mesoscopic systems

20.7.1 Shot noise in open mesoscopic systems

Shot noise in quantum point contacts. The suppression of the shot noise in mesoscopic conductors below the classical Schottky value can, for example, be observed in quantum point contacts. Figure 20.11 shows the comparison of the calculated and the measured shot noise in such a system. The measurement was performed at a temperature of 0.4 K

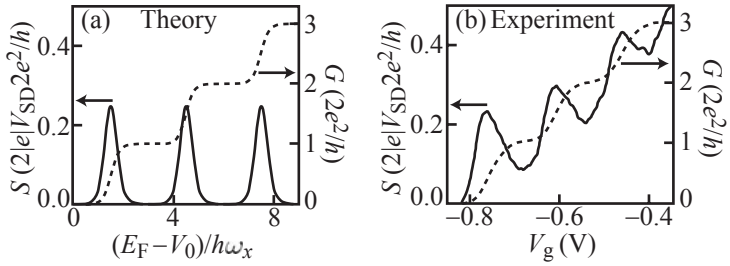


Fig. 20.11 (a) Shot noise calculated for a quantum point contact (solid line) in comparison to the conductance (dashed). For the calculation, the saddle potential model with $\omega_y = 3\omega_x$ was used. (De Jong and Beenakker, 1997. With kind permission of Springer Science and Business Media.) (b) Measured shot noise of a quantum point contact (solid line) in comparison to the measured conductance (dashed) at a temperature of $T = 0.4$ K. Courtesy of M. Reznikov and M. Heiblum, data similar to Reznikov *et al.*, 1995.

(similar to Reznikov *et al.* 1995). Similar experiments were, for example, performed in Kumar *et al.*, 1996, and Liu *et al.*, 1996. It can be seen that the conductance (dashed line) increases in the step-like fashion discussed before. The shot noise (solid line) oscillates with minima appearing at gate voltages where the conductance shows plateaus, and maxima where the conductance increase is steepest between two quantized steps. However, in the measurement, the shot noise is not completely suppressed on plateaus, in contrast to the theoretical calculation, and the shot noise peaks appear to be broader, showing that the saddle point approximation is not adequate for the description of the second and third conductance step in this example.

Shot noise in chaotic quantum billiards. Shot noise was also studied in coherent ballistic quantum structures with chaotic classical dynamics, so-called chaotic quantum billiards (Oberholzer *et al.*, 2002). The key significance of these experiments is that they give evidence for the quantum nature of scattering in these structures. The reason is that a quantum wave packet entering a chaotic billiard will follow a classical trajectory, but at the same time the wave packet will spread. If the dwell time τ_{dwell} of the wave packet in the billiard is larger than the Ehrenfest time

$$\tau_{\mathrm{E}} = \alpha^{-1} \ln \frac{L}{\lambda_{\mathrm{F}}},$$

where L is the characteristic size of the billiard, and α is a quantity related to the classical chaotic dynamics, then the wave packet has completely spread over the size of the billiard, and the correspondence principle between the quantum dynamics of the wave packet and the classical trajectory breaks down. The Fano factor has been predicted to follow (Agam *et al.*, 2000)

$$F = \frac{1}{4} \exp(-\tau_{\mathrm{E}}/\tau_{\mathrm{dwell}})$$

giving a Fano factor $F = 1/4$ in the extreme quantum limit $\tau_{\mathrm{dwell}} \gg \tau_{\mathrm{E}}$. This value was observed in Oberholzer *et al.*, 2002.

Shot noise in diffusive metallic wires. Another particularity of quantum shot noise arises in mesoscopic diffusive metallic wires. In contrast to macroscopic metallic systems which show no shot noise, coherent mesoscopic wires do. The reason is that in macroscopic systems of characteristic size L the shot noise of $L/l_\varphi \gg 1$ independently fluctuating segments averages out, whereas in systems with $L < l_\varphi$ this self-averaging is ineffective. Measurements of mesoscopic diffusive wires give a Fano factor $F = 1/3$ (Steinbach *et al.*, 1996; Henny *et al.*, 1999) as predicted by theory (Beenakker and Buttiker, 1992; Nagaev, 1992). The reason for the suppressed shot noise is the fact that the probability distribution of the $\mathcal{T}_n$ in eq. (20.22) in diffusive system is $p(\mathcal{T}_n) \propto \mathcal{T}^{-1}(1-\mathcal{T})^{-1/2}$. This is a surprising result, because it predicts large probabilities for closed channels with very small $\mathcal{T}_n$, large probabilities for open channels with $\mathcal{T}_n$ close to one, and only a very small probability for intermediate values of the transmission. The transmission channels therefore separate in a family of closed channels and a family of open channels. The significant fraction of open transmission channels leads to the Fano factor of 1/3.

Composite fermion charge from shot noise experiments. The classical shot noise formula of Schottky, eq. (20.17), predicts a proportionality between the shot noise power and the charge of the carriers responsible for electronic transport. Theory predicted a fractional charge for excitations in fractional quantum Hall minima. It was therefore an experimental challenge to provide experimental evidence for the fractional charge of quasiparticles via shot noise measurements. In these experiments, split gates are used to bring counterpropagating fractional edge channels, e.g., at filling factor $\nu = 1/3$ into tunneling distance. If the charge transfer is given by $|e|/3$, the charge of the Laughlin's fractional quasiparticles, the shot noise should be suppressed compared to the case where a charge $|e|$ is transferred. This suppression was indeed observed in experiments (Saminadayar *et al.*, 1997; de Picciotto *et al.*, 1997).

20.7.2 Shot noise and full counting statistics in quantum dots

Shot noise measurements in systems with very small transmission, such as quantum dots, are even more challenging than open systems, due to the small value of the transmission. Measurements based on conventional techniques were performed in Birk *et al.*, 1995. However, a very precise and new route of shot noise measurements on quantum dots uses the capability of quantum point contacts capacitively coupled to the quantum dot to detect the tunneling of individual electrons into and out of the dot.

Time-resolved noise measurements with quantum point contact charge detectors started with tuning up the bandwidth of conventional low-frequency setups (Schleser *et al.*, 2004). Using such setups, bandwidths

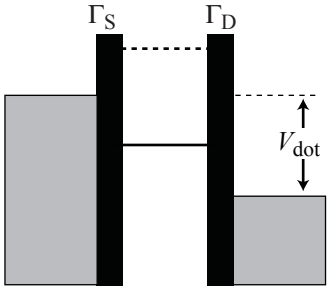


Fig. 20.12 Schematic energy diagram of the quantum dot showing the situation $V_{dot} \gg k_B T$.

of up to 30–40 kHz have been reported in the literature (Elzerman *et al.*, 2004; Gustavsson *et al.*, 2006), limiting the time resolution to the order of ten microseconds.

At fixed gate voltages, the charge detector witnesses electrons tunneling into and out of the quantum dot in real time. This manifests itself in random switching of the detector conductance between two distinct levels, called random telegraph noise, as shown in Fig. 18.5. When the conductance switches downwards, an electron has entered the dot, if it switches upwards, an electron has left the dot. If the quantum dot is in the single-level transport regime, the time-separations between tunneling-in and tunneling-out events follow the exponential decay laws

$$p_{\mathrm{in/out}}(t)dt = \Gamma_{\mathrm{in/out}} e^{-\Gamma_{\mathrm{in/out}} t} dt$$

with characteristic tunneling-in and tunneling-out rates $\Gamma_{\mathrm{in/out}}$. This decay law has been confirmed experimentally (Schleser *et al.*, 2004; Gustavsson *et al.*, 2006; MacLean *et al.*, 2007).

In the shot noise regime, where the source drain voltage $V_{dot} \gg k_B T$, but only a single quantum state is in the bias window (see Fig. 20.12), the rates $\Gamma_{\mathrm{in/out}}$ obtained from a time trace can be interpreted directly as the tunneling rates Γ_S and Γ_D (Gustavsson *et al.*, 2006). In this case, the electron tunneling into the dot will always originate from the source contact, and it will always tunnel out to the drain.

On the next level, correlations between subsequent tunneling-in and tunneling-out events at $V_{dot} \gg k_B T$ can be considered. For example, if we assume that such pairs of subsequent in/out events (in the following we call the pair an *event* for simplicity) are statistically independent, we find the statistical distribution

$$p_e(t)dt = dt \int_0^t dt' p_{\mathrm{in}}(t') p_{\mathrm{out}}(t - t') = \frac{\Gamma_{\mathrm{in}}\Gamma_{\mathrm{out}}}{\Gamma_{\mathrm{in}} - \Gamma_{\mathrm{out}}} \left(e^{-\Gamma_{\mathrm{in}} t} - e^{-\Gamma_{\mathrm{out}} t} \right) dt.$$

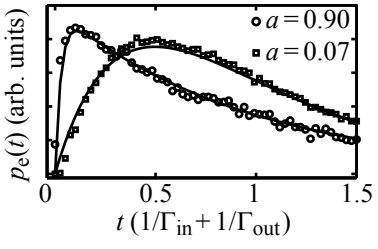


Fig. 20.13 Distribution $p_e(t)$ of times needed for one electron to traverse the quantum dot. Symbols are measured data points, solid lines are predictions of theory. The two distributions corresponding to different coupling asymmetries a (see text) are plotted on different vertical scales for clarity.

Figure 20.13 shows measurements of this distribution function for two different coupling asymmetries $a = (\Gamma_{\mathrm{in}} - \Gamma_{\mathrm{out}})/(\Gamma_{\mathrm{in}} + \Gamma_{\mathrm{out}})$. For almost symmetric coupling ($a = 0.07$) of the dot to the source and drain lead, there is a pronounced suppression of the distribution for small times. This is a direct consequence of the correlation between subsequently tunneling electrons brought about by the Coulomb blockade effect. The second electron has to wait with tunneling in until the first electron has tunneled out of the dot. This suppression becomes narrower in time for strongly asymmetric coupling ($a = 0.90$), because the system approaches the limit of a single barrier device in which no Coulomb blockade exists.

An alternative way of analyzing time-resolved single-electron tunneling traces, such as that shown in Fig. 18.5, is to look at full counting statistics. In order to do this analysis, a time trace of length T is divided into a reasonably large number of shorter segments of equal length ΔT. A histogram is then plotted for the distribution of the

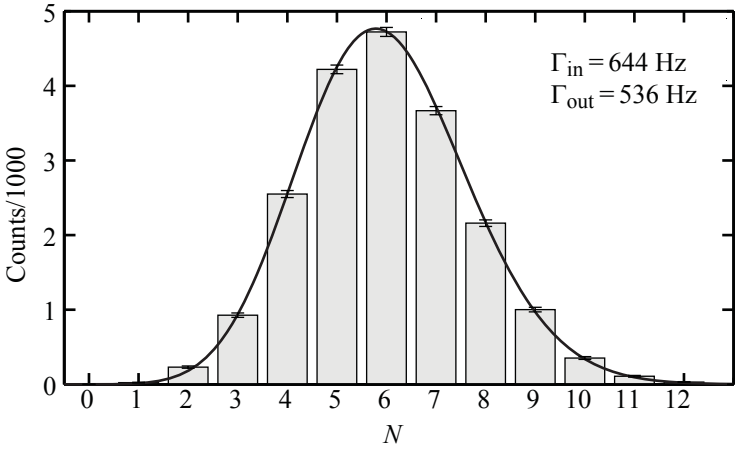


Fig. 20.14 Histogram of the full counting statistics of a quantum dot. The solid line is the theoretical prediction for the given rates $\Gamma_{\rm in}$ and $\Gamma_{\rm out}$.

number N of events found in the segments (an event is, for example, a down-switch of $I_{\rm QPC}$). An example of such a histogram, similar to those reported in Gustavsson *et al.*, 2006, is shown in Fig. 20.14. The mean value (first moment) $\langle N\rangle$ calculated with this histogram gives the mean current $I_{\rm dot} = e\langle N\rangle/\Delta T$ through the quantum dot. However, the width of the histogram, characterized by its second central moment (variance) $\langle (N-\langle N\rangle)^2\rangle$ is a measure of the fluctuations $\langle \Delta I^2\rangle = e^2\langle (N-\langle N\rangle)^2\rangle/\Delta T$ of the quantum dot current, meaning its shot noise. The shot noise for quantum dots has been calculated in Davies *et al.*, 1992, and later discussed in the framework of full counting statistics (Bagrets and Nazarov, 2003). While the shot noise of a single barrier device is expected to follow poissonian statistics with $\langle N\rangle = \langle (N-\langle N\rangle)^2\rangle$ (the mean equals the variance), for quantum dots the shot noise is expected to be suppressed as a result of the Coulomb-interaction-mediated correlations between tunneling electrons (see also the suppression of the distribution in Fig. 20.13 at short times). From Fig. 20.14 a variance $\langle (N-\langle N\rangle)^2\rangle \approx 3$ can be estimated, compared to a mean $\langle N\rangle \approx 6$, implying a reduction of the width by the Fano factor $F = 1/2$ compared to the poissonian case. Given the histogram shown in Fig. 20.14, even higher central moments, such as the skewness (3rd central moment) or the kurtosis (4th central moment) can be experimentally determined.

The full counting statistics can be found theoretically from a master equation approach. For example, in the single-level transport regime, the quantum dot system may be described by a two-state system with state 0 denoting zero, state 1 denoting one excess electron in the dot. We measure the current by counting the number N of electrons that transmit through the dot–drain barrier. We consider the case $V_{\rm dot} \gg k_{\rm B}T$ as depicted in Fig. 20.12(c), such that tunneling-in is only possible from the source (rate $\Gamma_{\rm S}$), and tunneling-out only through the drain (rate

Γ_D). The master equation is then given by

$$\begin{aligned} dp_0(t|N)/dt &= -\Gamma_S p_0(t|N) + \Gamma_D p_1(t|N-1) \\ dp_1(t|N)/dt &= -\Gamma_D p_1(t|N) + \Gamma_S p_0(t|N). \end{aligned}$$

Here, $p_n(t|N)$ is the probability that at time t, the system is found in state n, given that N electrons have been transferred into the drain lead since $t = 0$. At $t = 0$ we have the initial conditions $p_0(t = 0|N = 0) = 1$ and $p_n(t = 0|N \neq 0) = 0$. The rate equation can be solved using the discrete Fourier transform $p_n(t|\chi) = \sum_N p_n(t|N) \exp(iN\chi)$, where χ is called the counting field. We find the linear differential equation

$$\frac{d}{dt} \begin{pmatrix} p_0(t|\chi) \\ p_1(t|\chi) \end{pmatrix} = \begin{pmatrix} -\Gamma_S & \Gamma_D e^{i\chi} \\ \Gamma_S & -\Gamma_D \end{pmatrix} \begin{pmatrix} p_0(t|\chi) \\ p_1(t|\chi) \end{pmatrix},$$

which has the general solution $p_n(t|\chi) = \sum_{j=0}^{1} c_{nj} \exp[\lambda_j(\chi)t]$ with the $\lambda_j(\chi)$ being the eigenvalues of the coefficient matrix. For times t large compared to the correlation time $(\Gamma_S + \Gamma_D)^{-1}$ (Machlup, 1954), the solution is governed by the eigenvalue with the smallest negative real part (say, λ_0) giving the slowest decay. The full counting statistics, i.e., the probability that N electrons have been transferred through the dot after time ΔT is given by

$$P_N(\Delta T) = \sum_{n=0}^{1} p_n(\Delta T|N) = \frac{1}{2\pi} \int d\chi e^{-iN\chi} \sum_{n=0}^{1} p_n(\Delta T|\chi).$$

The logarithm of its Fourier transform is the cumulant generating function $S(\chi)$, which has the large ΔT limit $S_{\Delta T}(\chi) = \lambda_0(\chi)\Delta T$. The mean current is given by the first cumulant $\langle N \rangle = -idS/d\chi|_{\chi=0}$, and the shot noise by the second cumulant $\langle (N - \langle N \rangle)^2 \rangle = -d^2S/d\chi^2|_{\chi=0}$. The resulting full counting statistics, which has been worked out in Bagrets and Nazarov, 2003, is plotted as a solid line in Fig. 20.14, and shows excellent agreement with the measured histogram. Finite bandwidth corrections (Naaman and Aumentado, 2006) have been taken into account. More details about the analysis of full counting statistics data can be found in the review by Gustavsson *et al.*, 2007, and in the overview article Gustavsson *et al.*, 2008.

Further reading

- Easy reading: Beenakker and Schonenberger 2003; Belzig 2005.
- Reviews: De Jong and Beenakker 1997; Blanter and Buttiker 2000; Martin 2005.
- Book: Nazarov 2003.

Exercises

(20.1) You tune a quantum point contact in GaAs at the temperature $T = 1.7\,\mathrm{K}$ to the conductance $G = e^2/h$, i.e., to the transition between complete pinch-off and the first conductance plateau.

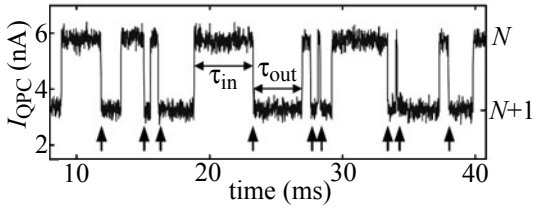
(a) Calculate the spectral density of the thermal noise.

(b) What is the lowest voltage that you would have to apply at least to the quantum point contact in order to have the shot noise dominate over the thermal noise (assume the transmission to be energy-independent)? What is the current through the point contact under these conditions?

(c) Compare the spectral density of the thermal noise of the point contact with that of the $5\,\mathrm{k}\Omega$ ohmic contacts through which the point contact is connected to the external circuit.

(d) Discuss what further noise sources you have to consider if you intend to measure the noise of the point contact.

(20.2) Consider a quantum dot coupled to a quantum point contact charge detector. The quantum dot is very weakly coupled to its source and drain leads such that electrons tunneling into and out of the quantum dot can be counted in real time using the quantum point contact. With a source–drain voltage V_{SD} applied to the quantum dot ($|e|V_{\mathrm{SD}} \gg k_{\mathrm{B}}T$), but only a single energy level in the transport window ($|e|V_{\mathrm{SD}} \gg \Delta$) the following measurement of the quantum point contact current was made:



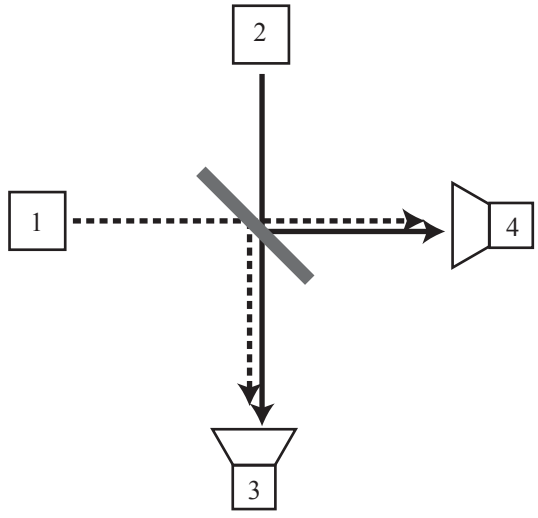
(a) What is the statistical probability density distribution for the times τ_{in} and τ_{out}?

(b) Estimate the tunneling rates Γ_{S} and Γ_{D} from the data.

(c) What is the time-averaged occupation probability of the energy level in the dot, given the data?

(d) Estimate the average tunneling current through the quantum dot from the data.

(20.3) From shot noise measurements, information about correlations between particles can be obtained. In this exercise a scattering experiment with a half-transparent beam splitter is investigated, such as that realized if a single quantum Hall edge channel impinges onto a quantum point contact tuned to transmission 1/2. The schematic setup shown in the figure consists of two particle sources 1 and 2, a beam splitter and two detectors 3 and 4. A particle emitted by source 1 or 2 will be transmitted with equal probability to detector 3 or 4.

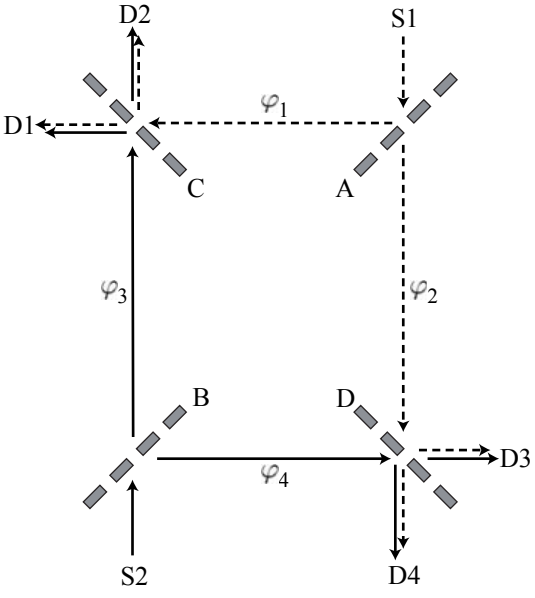


If two identical particles are emitted simultaneously by the two sources, we have to treat a two-particle scattering problem. Two-particle fermion- and boson states will have to obey different symmetry upon particle exchange. We look at the probabilities $p(1,1)$ of detecting an event in both detectors, $p(2,0)$ of two events in detector 3, or $p(0,2)$ of two events in detector 4.

(a) Show that for fermions (electrons) $p(2,0) = 0$, $p(1,1) = 1$, and $p(0,2) = 0$.

(b) Show that, in contrast, for bosons (e.g., photons), $p(2,0) = 1/2$, $p(1,1) = 0$, and $p(0,2) = 1/2$.

(20.4) Noise measurements provide more information about a mesoscopic system than measurements of the current. This can be impressively seen in the experiment Neder *et al.*, 2007*b*. In this problem you will show that current measurements in the geometry investigated in this paper do not show an Aharonov–Bohm interference effect, whereas the interference can be seen in noise correlation measurements on two contacts of the structure. A schematic of the investigated structure is depicted below.



Electrons are injected from two independent ohmic contacts S1 and S2 into one-dimensional quantum Hall edge channels. Quantum point contacts (QPCs) A, B, C, D act as beam splitters for the incoming electron wave. Assume that incoming amplitudes (A_1, A_2) and outgoing amplitudes (B_1, B_2) at each beam splitter are related by the scattering matrix

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix},$$

which is assumed to be the same for all four QPCs. Between the beam splitters, the electron wave function accumulates phases φ_1, φ_2, φ_3, and φ_4, as indicated in the figure. Current can be measured using ohmic contacts D1, D2, D3, D4.

(a) Find reasons why no Aharonov–Bohm effect can be observed in the current at D1, D2, D3, or D4, if electrons are only injected from S1 into the structure.

(b) Write down the expressions for the two outgoing single-particle wave functions at the four detector contacts D1, D2, D3, D4, assuming that single electrons are injected from either S1 or from S2.

(c) Form the correctly antisymmetrized fermionic wave function for two electrons injected simultaneously at S1 and S2.

(d) How does the probability of detecting correlated electrons at D2 and D4 depend on the four phases φ_1, φ_2, φ_3, and φ_4?

(e) Find reasons why a measurement of the correlated noise measured between contacts D2 and D4 would be sensitive to small changes of the magnetic field via the Aharonov–Bohm phase.

Interference effects in nanostructures II

21

21.1 The Fano effect

The Fano effect is a phenomenon that arises in many areas of physics. It is caused by the interference of a resonant coherent scattering channel of a particle with a nonresonant continuum channel. We will discuss the meaning of this general description below. The more familiar resonant scattering phenomenon was described in Breit and Wigner, 1936, considering neutrons scattering at atoms, and the interference of a Breit–Wigner resonance with a nonresonant scattering channel was discussed later by Feshbach, Peaslee, and Weisskopf (Feshbach *et al.*, 1947). They found that interference with the nonresonant scattering channel leads to a characteristic change in the resonance lineshape, which was described in 1961 by Ugo Fano in connection with resonant scattering of photons and atoms (Fano, 1961). It has to be mentioned, however, that Fano had already presented his theory leading to the asymmetric lineshapes in a paper published in 1935 (Fano, 1935).

In experiments on semiconductors the Fano effect arises, for example, in Raman scattering (Cerdeira *et al.*, 1973), in optical absorption in quantum wells (Maschke *et al.*, 1991; Faist *et al.*, 1997), and in mesoscopic transport through one-dimensional channels with resonances (see, e.g., Chu and Sorbello, 1989; McEuen *et al.*, 1990). A very comprehensive theoretical description can be found in Nockel and Stone, 1994.

One of the simplest geometries in which the Fano effect has been observed in mesoscopic transport is a quantum dot coherently side-coupled to a one-dimensional channel (quantum point contact). Such a structure based on a two-dimensional electron gas in a Ga[Al]As heterostructure is shown in the inset of Fig. 21.1. If the conductance of the point contact is measured as a function of the plunger gate voltage V_g by applying a voltage between source and drain, the resonances of the quantum dot can be seen as shown in Fig. 21.1. The line shape of these resonances is very different from the sharp symmetric peaks usually observed in the Coulomb blockade regime.

The essence of the Fano effect in this structure can be discussed without considering the Coulomb interaction among electrons in the structure. Since the quantum dot has no 'exit', but is only coupled via one single opening to the channel, we consider the one-dimensional total reflection at a resonator in the picture of interfering Feynman paths as

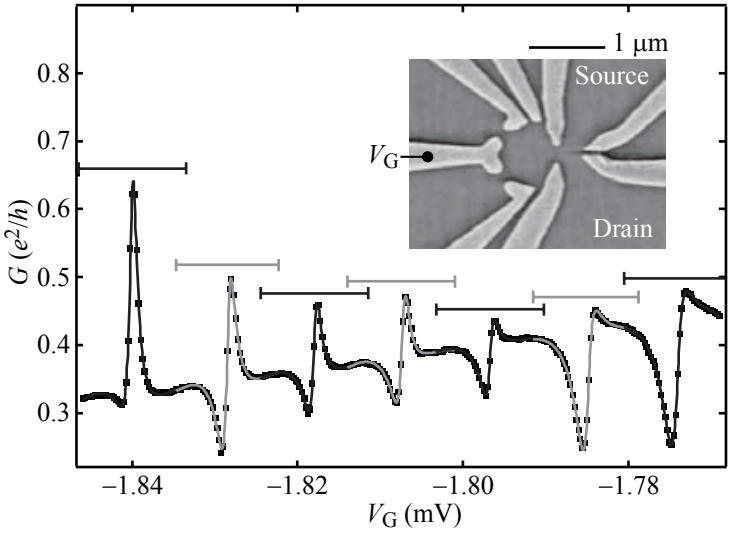


Fig. 21.1 Fano resonances in the conductance of a quantum point contact with side-coupled quantum dot, measured as a function of the plunger gate voltage V_g. Inset: Structure showing the arrangement of gate electrodes on the sample surface which allows the formation of the conducting channel with side-coupled quantum dot. (Reprinted with permission from Johnson *et al.*, 2004. Copyright 2004 by the American Physical Society.)

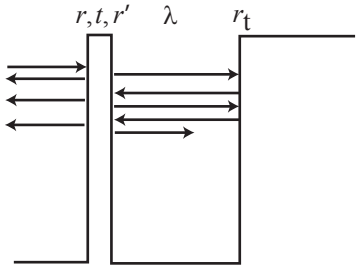


Fig. 21.2 Schematic illustration showing a square potential minimum coupled to a continuum of states to the left via a thin tunneling barrier with reflection coefficients r and r', and transmission coefficient t. Propagation within the potential well leads to phase accumulation described by λ. At the potential step to the right, particles are totally reflected as described by the reflection amplitude r_t.

schematically shown in Fig. 21.2. The reflected amplitude can be written as

$$r_{\rm res} = r + t\lambda r_{\rm t}\lambda \left[\sum_{n=0}^{\infty}(r'\lambda r_{\rm t}\lambda)^n\right] t = r + \frac{t\lambda r_{\rm t}\lambda t}{1 - r'\lambda r_{\rm t}\lambda}.$$

Here λ describes the propagation between the barrier connecting the dot with the channel and the barrier where total reflection takes place. The quantity $r_{\rm t}$ is the amplitude of total reflection, and r, r', and t are the reflection and transmission coefficients of the barrier. For simplicity we consider only this one-dimensional resonator and write the amplitudes as

$$r = \sqrt{R}e^{i\alpha}, \quad r' = -\sqrt{R}e^{i\gamma}, \quad t = \sqrt{1-R}e^{i(\alpha+\gamma)/2},$$
$$r_{\rm t} = e^{i\delta}, \qquad \lambda = e^{i\beta}.$$

Here, the α, β, γ, and δ are phases and R is the reflection probability of the barrier coupling the dot to the channel. The larger R is, the weaker is the dot coupled to the channel. Introducing these amplitudes in the Feynman path result, we obtain

$$r_{\rm res} = e^{i\alpha}\frac{\sqrt{R} + e^{i\theta_{\rm D}}}{1 + \sqrt{R}e^{i\theta_{\rm D}}},$$

where $\Theta_{\rm D} = \gamma + 2\beta + \delta$ is the phase accumulated by an electron on a round trip between the barriers. It is straightforward to show that the magnitude of the numerator and of the denominator in this expression is the same, i.e., $|r_{\rm res}|^2 = 1$, as expected for total reflection. We are therefore only interested in the phase angle $\varphi = \arg(r_{\rm res})$. In order to determine φ, we introduce the coupling parameter $\gamma := (1 - \sqrt{R})/(1 + \sqrt{R})$ (i.e., $\sqrt{R} = (1-\gamma)/(1+\gamma)$) and find

$$r_{\rm res} = e^{i\alpha}\frac{\cos(\theta_{\rm D}/2) + i\gamma\sin(\theta_{\rm D}/2)}{\cos(\theta_{\rm D}/2) - i\gamma\sin(\theta_{\rm D}/2)}.$$

From this expression, we read

$$\tan[(\varphi-\alpha)/2]=\gamma\tan(\theta_{\rm D}/2),$$

which creates the nonlinear map between φ and $\theta_{\rm D}$ shown in Fig. 21.3. Resonances occur for $\theta_{\rm D}=(2p+1)\pi\equiv\theta_p$, where p is an integer. Near resonance the numerator and denominator of $r_{\rm res}$ can be expanded to first order in $\theta_{\rm D}-\theta_p$ giving

$$r_{\rm res}=-e^{i\alpha}\frac{1+i(\theta-\theta_p)/2\gamma}{1-i(\theta-\theta_p)/2\gamma}.$$

The relation to energy is established via

$$\frac{\theta-\theta_p}{2\gamma}=\frac{1}{2\gamma}\left.\frac{d\theta(E)}{dE}\right|_{E=E_p}(E-E_p)=\frac{E-E_p}{\Gamma_p}=\epsilon$$

with

$$\frac{1}{\Gamma_p}=\frac{1}{2\gamma}\left.\frac{d\theta(E)}{dE}\right|_{E=E_p}.$$

As a result we obtain, near resonance,

$$r_{\rm res}=-e^{i\alpha}\frac{1+i\epsilon}{1-i\epsilon}=e^{i\varphi},\ \text{ with }\ \varphi=\alpha+\pi+2\vartheta,\ \text{ where }\ \tan(\vartheta)=\epsilon.$$

Returning to the problem of the cavity coupled to the channel, we can see that there are the two alternative paths depicted in Fig. 21.4: either the electron is directly transmitted through the channel, or it also visits the quantum dot during the course of the transmission. In the Feynman path description of the total transmission coefficient we can write this as

$$t_{\rm W}=t_{\rm d}+b_1r_{\rm res}b_2.$$

The amplitude $t_{\rm d}$ describes the direct transmission from source to drain, b_1 is the amplitude for the transmission from the source contact into the dot, and b_2 that for transmission from the dot into the drain.

For the calculation of the transmission probability of the wire close to a resonance of the dot we write

$$t_{\rm W}=|t_{\rm d}|+f_0e^{2i(\phi_{\rm F}+\vartheta)},$$

where $f_0=|b_1||b_2|$, $2\phi_{\rm F}=\alpha+\pi-\arg(t_{\rm d})$, and we have omitted an irrelevant overall phase factor. The angle $2\phi_{\rm F}$ can be seen as the phase difference between the resonant and the nonresonant path excluding the resonant behavior of the dot.

As in conventional two-path interference problems the transmission probability is then

$$\begin{aligned}T_{\rm W}&=|t_{\rm W}|^2=t_{\rm d}^2+f_0^2+2t_{\rm d}f_0\cos[2(\phi_{\rm F}+\vartheta)]\\&=(t_{\rm d}-f_0)^2+2t_{\rm d}f_0\left\{1+\cos[2(\phi_{\rm F}+\vartheta)]\right\}\\&=(t_{\rm d}-f_0)^2+4t_{\rm d}f_0\cos^2(\phi_{\rm F}+\vartheta)\\&=(t_{\rm d}-f_0)^2+\frac{4t_{\rm d}f_0}{1+\tan^2(\phi_{\rm F}+\vartheta)}.\end{aligned}$$

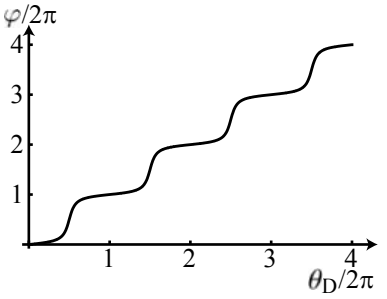


Fig. 21.3 Phase φ of the reflection coefficient of a cavity as a function of the round trip phase $\theta_{\rm D}$ within the cavity, assuming $\alpha=0$.

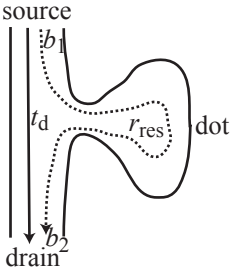


Fig. 21.4 Schematic illustration showing the interference of two alternative paths: the direct transmission with amplitude $t_{\rm d}$ and the resonant transmission with amplitue $b_1r_{\rm res}b_2$.

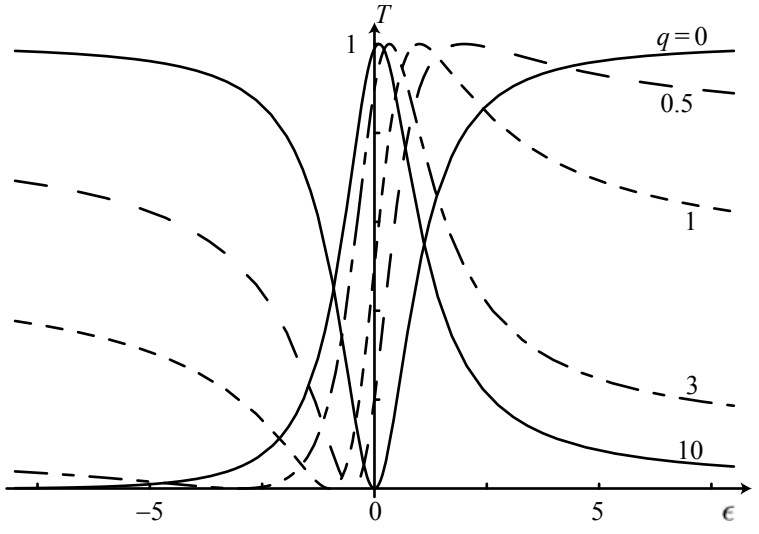


Fig. 21.5 Linie shapes of Fano resonances for different Fano parameters q.

Further application of addition theorems and some algebra lead to the relation

$$\frac{1}{1+\tan^2(\varphi_\mathrm{F}+\vartheta)} = \frac{(-\cot\phi_\mathrm{F}+\tan\vartheta)^2}{(1+\cot^2\phi_\mathrm{F})(1+\tan^2\vartheta)}.$$

We now define the so-called *Fano parameter* $q = -\cot\phi_\mathrm{F} = \tan[(\alpha - \arg(t_\mathrm{d}))/2]$, insert $\tan\vartheta = \epsilon$ and obtain the Fano formula

$$T_\mathrm{W} = (t_\mathrm{d} - f_0)^2 + \frac{4t_\mathrm{d}f_0}{1+q^2}\frac{(q+\epsilon)^2}{1+\epsilon^2}. \tag{21.1}$$

The shape of the transmission resonance depends on the Fano parameter q which describes the phase difference between the direct and the resonant path. Figure 21.5 shows line shapes for different Fano parameters assuming that $t_\mathrm{d} = f_0 = 1/2$. For $q = 0$ an antiresonance is seen with completely suppressed transmission on resonance. If $q \to \pm\infty$ a Breit–Wigner resonance with transmission one on resonance is recovered. For $q = \pm 1$ a completely asymmetric lineshape results. In this case, the direct and the resonant path interfere constructively on one side and destructively on the other side of the resonance.

It can be shown that in systems with time-reversal symmetry (like our example), the Fano parameter q is a real number (Nockel and Stone, 1994; Clerk *et al.*, 2001). For real q the probability T vanishes for $\epsilon = -q$.

Fano resonances can also arise in the transmission through a single quantum dot, if the tunneling coupling to the leads is strong. It turns out that different quantum dot states can couple with very different strengths to the leads. The resonant path would then be the transmission through a weakly coupled level of the quantum dot which is sharply defined in energy. The nonresonant path is the transmission through a very strongly coupled and therefore energetically strongly broadened quantum dot level which acts as a 'continuum'. Figure 21.6 shows such resonances in the conductance through a quantum dot.

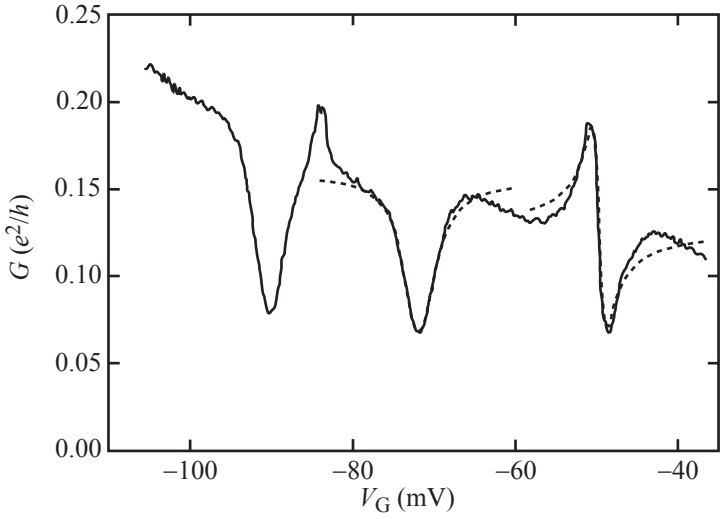


Fig. 21.6 Fano resonances in the conductance through a quantum dot. (Reprinted with permission from Gores *et al.*, 2000. Copyright 2000 by the American Physical Society.)

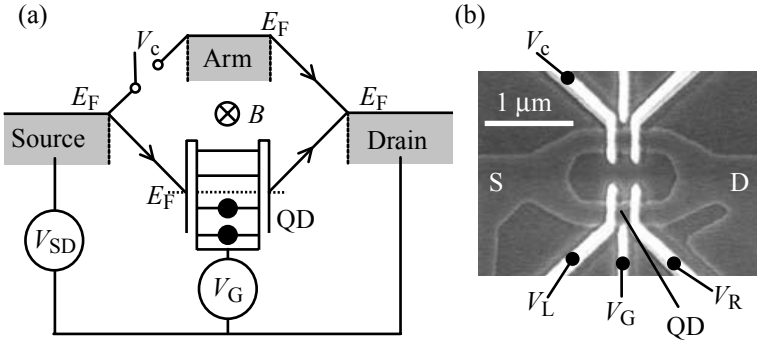


Fig. 21.7 (a) Schematic illustration of the experimental setup where an electron injected from the source contact interferes via the two alternative paths 'Arm' and 'QD'. (b) The scanning electron micrograph of the sample surface shows the arrangement of the top gates and the edges of the wet-chemically etched mesa edges. (Reprinted with permission from Kobayashi *et al.*, 2002. Copyright 2002 by the American Physical Society.)

Another variant of the Fano effect is observed in quantum ring structures, where a quantum dot is embedded in one arm of the ring. Figure 21.7 shows such a structure. The Fano effect arises if the Breit–Wigner resonance in the transmission through the quantum dot interferes with the direct transmission through the second arm (reference arm) of the interferometer. In the experiment shown here, it was possible to switch the transmission through the reference arm on and off using a gate electrode. The measurement of the resonances in these two cases (Fig. 21.8) shows how the additional interference with the transmission through the reference arm acts on the resonance lineshape.

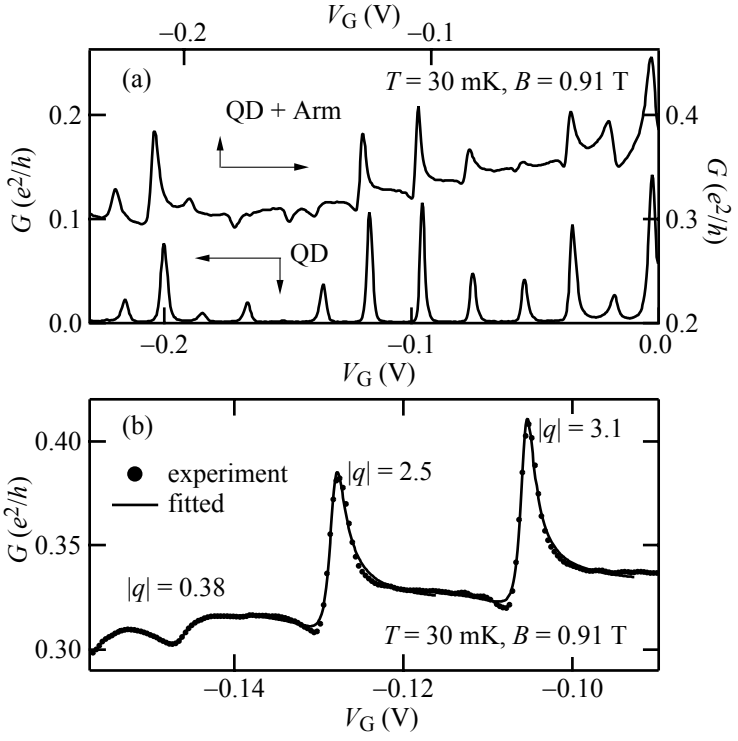


Fig. 21.8 (a) Comparison between conductances of the system with and without reference arm.When the reference arm transmits, the previous Coulomb-blockade resonances develop into Fano resonances. (b) Fits of the Fano-line shape to the measured data, and corresponding Fano parameters. (Reprinted with permission from Kobayashi *et al.*, 2002. Copyright 2002 by the American Physical Society.)

21.2 Measurements of the transmission phase

Basic concept. The measurement of the conductance of a mesoscopic system is, following the spirit of the Landauer–Büttiker formalism, a measurement of the transmission probability $\mathcal{T}$. We can measure this quantity as a function of some system parameter, such as a gate voltage. The probability $\mathcal{T}$ is given by the squared magnitude of a complex-valued probability amplitude t, i.e., $\mathcal{T} = |t|^2$. The probability amplitude t can be expressed by a pair of numbers, for example, by the magnitude $a = |t|$ and the phase $\theta = \arg(t)$, each having its own dependence on the system parameter. The conductance measured, for example, as a function of the gate voltage, can be seen as a measurement of the gate voltage dependence of a, but that of the phase θ can usually not be retrieved. It is therefore of interest to design experiments in which the dependence of θ on some system parameter can also be measured, because this would yield additional information about the system under investigation.

From previous discussions of interference phenomena we have learned that it is only the relative phase, i.e., the difference of the phases of two alternative processes 1 and 2 that comes into play in interference [cf., eq.(14.1)]. One could therefore try to measure the voltage (V_G) dependence (say) of the transmission amplitude *and* phase through a

mesoscopic system in an interference experiment in which only one of the two alternative transport paths depends on the voltage while the other is voltage independent. On a conceptual level, such a measurement would give

$$\mathcal{T}(V_\mathrm{G}) = |t_1(V_\mathrm{G}) + t_2|^2 = a_1^2(V_\mathrm{G}) + a_2^2 + 2a_1(V_\mathrm{G})a_2\cos[\theta_1(V_\mathrm{G}) - \theta_2].$$

We see that it is still not possible in general to separate the terms containing the transmission phase $\theta_1(V_\mathrm{G})$ and those containing the transmission amplitude $a_1(V_\mathrm{G})$ from a measurement of $\mathcal{T}(V_\mathrm{G})$.

We therefore need some second parameter which acts on the interference term *only* without affecting the amplitudes a_1 and a_2. What comes to our rescue here is the Aharonov–Bohm effect. If we can arrange the two alternative transmission paths in such a way that they enclose a magnetic flux ϕ, and if we can make sure that the magnetic field generating this flux does not influence the amplitudes a_1 and a_2, then we obtain the transmission

$$\mathcal{T}(V_\mathrm{G}, \phi) = a_1^2(V_\mathrm{G}) + a_2^2 + 2a_1(V_\mathrm{G})a_2\cos[\theta_1(V_\mathrm{G}) - \theta_2 + \phi/\phi_0], \quad (21.2)$$

with $\phi_0 = h/e$ being the magnetic flux quantum.

The concept introduced above was realized in Schuster *et al.*, 1997, in an attempt to measure the transmission phase of a quantum dot. The significance of this experiment is first of all that it gives evidence for *coherent* transmission through a quantum dot. In addition, it demonstrates that the concept of the phase measurement outlined above works in principle. The quantum dot is a suitable system for such an experiment, because the transmission amplitude a_1 through a dot of characteristic size L changes only at the magnetic field scale $\Delta B = \phi_0/L^2$. If the Aharonov–Bohm interferometer is made much larger than L, the Aharonov–Bohm oscillations will occur at a field scale much smaller than ΔB.

Figure 21.9 shows the sample used for the corresponding measurement. The metallic gates are used to deplete the two-dimensional electron gas residing below the surface. The quantum dot is embedded in one path of the two-path interferometer indicated by the arrows. In the experiment, a voltage is applied between the source contact and the grounded base regions. The collector voltage is measured. At the measurement temperature of 80 mK electrons travel ballistically and phase-coherently throughout the whole structure. The analysis in the framework of the Landauer–Büttiker theory leads to the collector voltage

$$V_\mathrm{c} = \frac{\mathcal{T}_\mathrm{cs}}{N_\mathrm{c} - \mathcal{R}_\mathrm{c}} V_\mathrm{s},$$

which is proportional to the source–collector transmission $\mathcal{T}_\mathrm{cs}$ given by eq. (21.2).

Before we look at the experimental results, we use eq. (18.28) to obtain an expectation for the magnitude a_1 and the phase θ_1 of the transmission through a quantum dot. Figure 21.10 shows the evolution of the

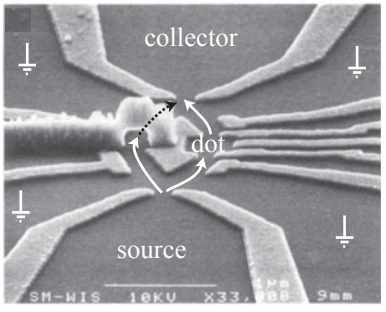


Fig. 21.9 Scanning electron micrograph of the sample used for the transmission phase measurement. A voltage is applied between the source contact and ground. The voltage between the collector contact and ground is measured. (Schuster *et al.*, 1997. Reprinted by permission from Macmillan Publishers Ltd, copyright 1997.)

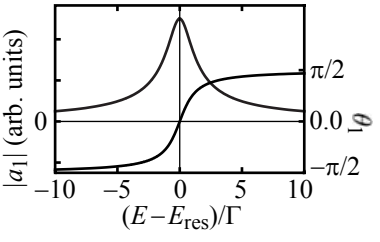


Fig. 21.10 Calculated magnitude and phase of the resonant transmission through a single energy level in the Lorentz approximation.

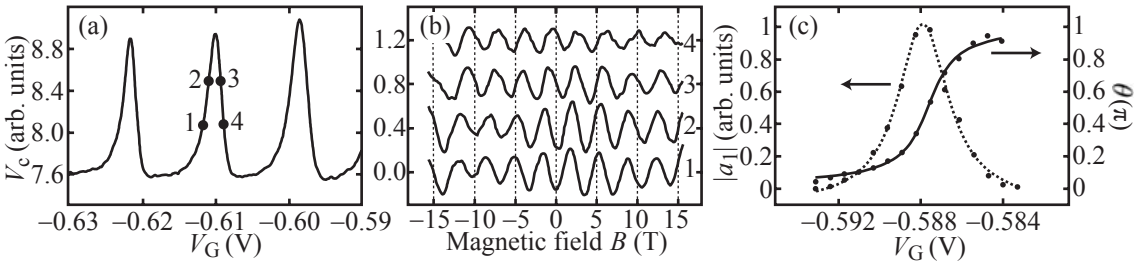


Fig. 21.11 (a) Collector voltage V_c measured as a function of the plunger gate voltage V_G of the quantum dot. (b) Aharonov–Bohm oscillations of V_c measured as a function of magnetic field. The Aharonov–Bohm period is indicated by vertical dashed lines. (c) Amplitude and phase extracted for one particular resonance. (Schuster *et al.*, 1997. Reprinted by permission from Macmillan Publishers Ltd, copyright 1997.)

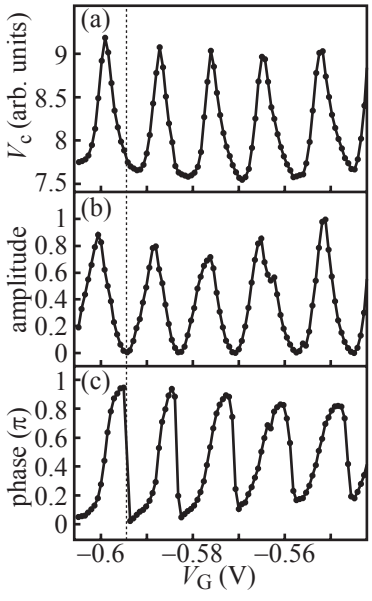


Fig. 21.12 (a) Collector voltage V_c at zero magnetic field as a function of the quantum dot plunger gate voltage V_G. (b) Amplitude of the Aharonov–Bohm oscillation as a function of V_G. (c) Phase of the Aharonov–Bohm oscillation at zero magnetic field as a function of V_G. (Schuster *et al.*, 1997. Reprinted by permission from Macmillan Publishers Ltd, copyright 1997.)

magnitude of the transmission, and the transmission phase as a function of the energy difference from resonance $E - E_{\rm res}$, normalized to the transmission broadening Γ. We see that the magnitude of the transmission amplitude resembles the peak-shaped structure that is also seen in the transmission probability in Fig. 18.30. The phase shows a gradual increase by π with the steepest slope at resonance. This is what we hope to observe in the experiment.

Experimental measurement of the transmission phase of a quantum dot. The measurement procedure will now be as follows: First the collector voltage is measured as a function of the plunger gate voltage V_G of the quantum dot as shown in Fig. 21.11(a). A series of conductance resonances are observed. Between the resonances the current does not drop to zero, because the reference arm is always open and gives rise to a finite collector voltage, even if the dot is completely Coulomb blockaded. Then measurements at fixed V_G deliver the ϕ-dependence of $\mathcal{T}$. We observe the oscillatory modulation due to the cosine interference term, as shown in Fig. 21.11(b) for the four different V_G indicated in (a). This allows us to determine the phase difference $\theta_1(V_G) - \theta_2$ at $\phi = 0$ on an absolute scale. Repeating such measurements for a set of gate voltages V_G, we can observe the evolution of $\theta_1(V_G)$ relative to θ_2, which is as close as we can get to the measurement of $\theta_1(V_G)$. The result of this procedure is depicted in Fig. 21.11(c). At the same time, the simultaneous measurement of the magnetic field averaged contribution to the conductance $|a_1|^1 + |a_2|^2$ and the Aharonov–Bohm oscillation amplitude $2|a_1||a_2|$ allows us to determine $|a_1(V_G)|$ which is also shown in (c). We see that the transmission amplitude follows a lorentzian line shape, and the phase shows a smooth increment of π as the resonance is traversed, as expected for a lorentzian resonance [cf., Fig. 21.10].

Figure 21.12 shows the result of this measurement procedure for a series of resonances. The increase of the transmission phase by π is consistently observed for each resonance. However, there are unexpected jumps of the phase by $-\pi$ between resonances, also called *phase lapses*, which have given rise to a lot of discussion in recent years about their

origin. In the simple model of lorentzian resonances, subsequent resonances should simply accumulate the phase. According to eq. (18.28), the transmission amplitudes of subsequent resonances have alternating sign, which means that their phases are shifted by π relative to each other. This results in a steady step-like increase of the transmission phase over many resonances, in contrast to the experimental results. It has been claimed that this type of observed behavior is universal for many-electron quantum dots. More recent experiments on few-electron quantum dots, however, have shown a steady increase of the phase without phase lapses (Avinun–Kalish *et al.*, 2005).

21.3 Controlled decoherence experiments

What are we aiming at? We discussed in section 14.7 how decoherence of quantum states comes about in principle. Usually decoherence in a quantum system is caused by some kind of environment with many degrees of freedom that are not well controlled. Fluctuations in the environment couple back into the system and lead to the randomization of the phase, a process that we call decoherence. We have seen that the thermal bath of electrons can lead to the decoherence of individual interfering electrons. Alternatively, the environment may be represented by a phonon system in the host crystal which gradually destroys phase coherence at increasing temperature, or by a photon bath coupling to the electrons. However, all these mechanisms can be greatly suppressed at low temperatures, i.e., if we do experiments in a dilution refrigerator below a temperature of 100 mK, say.

On the other hand, we have shown in previous chapters of this book, how well-controllable quantum systems can nowadays be fabricated and investigated. An obvious question to ask is therefore whether we can use advanced fabrication techniques to design experiments where the environment is tailored in such a way that its previously uncontrolled back-action can be controlled by experimental parameters.

This question touches fundamental issues of measurement in quantum mechanics. We all know that, in the paradigmatic double-slit interference experiment, the interference vanishes if we invent means to measure which of the two slits the particle took, i.e., if we extract *which-path* information (Feynman *et al.*, 2006). Uncontrolled decoherence caused by an environment can be interpreted as a measurement process that the environment performs on the system (although no experimentalist can ever retrieve the acquired information from the environment). As a consequence, the quest to control decoherence experimentally requires us to design well-controllable coupled interferometer–detector setups.

Throughout this book we have discussed a number of electronic interferometers with optical analogues. The first one was the Aharonov–Bohm interferometer in chapter 14, which can be seen as the analogue of the optical double slit interferometer. The second one was the electronic Mach–Zehnder interferometer in section 16.5, and the third was

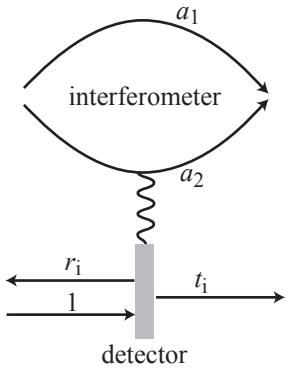


Fig. 21.13 Schematic interferometer–detector arrangement as it is considered in the text.

Table 21.1

system electron	detector electron
slit 1	reflected
slit 1	transmitted
slit 2	reflected
slit 2	transmitted

the quantum dot system which is the electronic version of the Fabry–Perot interferometer (section 18.3.1). How can we design which-path detectors for these interferometers that would allow us to study and control the process of decoherence?

Principle of operation of interferometer–detector arrangements. This question has been tackled by three experiments involving the three interferometer types discussed above, and using a quantum point contact detector for sensing charge capacitively (Buks *et al.*, 1998; Sprinzak *et al.*, 2000; Neder *et al.*, 2007*a*).

In all these experiments, the detector is a quantum point contact tuned to the transition region between complete pinch-off and the first plateau. The schematic setup of the interferometer–detector arrangement that we are interested in here is shown in Fig. 21.13. The interferometer allows an electron to pass through two alternative paths with probability amplitudes a_1 and a_2 that are brought back to interference at the exit of the interferometer. The detector consists of a potential barrier onto which detector electrons impinge. The electron may be either transmitted or reflected with amplitudes t_i or r_i respectively. The index $i \in \{1, 2\}$ makes the connection to the interferometer, because the transmission and reflection amplitudes may depend on whether the interferometer electron takes the upper or the lower path.

If we consider the injection of one electron in the interferometer interacting with one simultaneously injected electron in the detector, we have the four possible outcomes of the experiment listed in Table 21.1. In the spirit of section 14.7 we can write the entangled state between the detector and the interferometer as

$$\psi(x,\eta) = a_1\varphi_1(x)[r_1\chi_r(\eta) + t_1\chi_t(\eta)] + a_2\varphi_2(x)[r_2\chi_r(\eta) + t_2\chi_t(\eta)],$$

where the wave functions $\varphi_{1/2}$ describe the states of the system electron going through slit 1 or 2, respectively, and the wavefunctions $\chi_{\mathrm{r/t}}$ describe the states of the detector electron being reflected or transmitted. In the following we assume the two detector states to be orthogonal, i.e., perfectly distinguishable. The a_1, a_2, t_1, t_2, r_1, and r_2 are the corresponding probability amplitudes for the processes in the system and the detector. They obey the relations $|a_1|^2 + |a_2|^2 = 1$, $|t_1|^2 + |r_1|^2 = 1$, and $|t_2|^2 + |r_2|^2 = 1$. Taking the magnitude squared of this wave function and integrating out the detector variable η, we find

$$|\psi(x,\eta)|^2 = |a_1|^2|\varphi_1(x)|^2 + |a_2|^2|\varphi_2(x)|^2 + a_1a_2^\star\varphi_1(x)\varphi_2^\star(x)(r_1r_2^\star + t_1t_2^\star) + \text{c.c.}$$

Here the two terms to the right of the equal sign in the first line are the classical contributions, whereas the second line describes quantum interference. From this expression we identify the transmission probability $\mathcal{T}_\mathrm{i}$ through the interferometer to be

$$\mathcal{T}_\mathrm{i} = |a_1|^2 + |a_2|^2 + [a_1a_2^\star(r_1r_2^\star + t_1t_2^\star) + \text{c.c.}].$$

We see here that if $r_1 = r_2$, and $t_1 = t_2$, meaning that the detector cannot distinguish which of the two slits the system electron takes, the decoherence-free interference term is recovered. If this is not the case, the expression $r_1 r_2^\star + t_1 t_2^\star$ describes the decoherence caused by the electron in the quantum point contact detector (Averin and Sukhorukov, 2005). Decoherence can arise, if r_1 and r_2 (t_1 and t_2) differ in magnitude, or in phase, or in both. These are the ways in which the detector can in principle acquire information about the path that the interferometer electron took. We also note that the classical contribution to the transmission is not affected by the presence of the detector.

In the next step we introduce magnitude and phase of the amplitudes explicitly and write

$$\mathcal{T}_\mathrm{i} = |a_1|^2 + |a_2|^2 + |a_1||a_2|[e^{i(\varphi+\delta)}(|r_1||r_2|e^{i\Delta\theta_\mathrm{r}} + |t_1||t_2|e^{i\Delta\theta_\mathrm{t}}) + \mathrm{c.c.}],$$

where $\varphi = 2\pi\phi/\phi_0$ is the Aharonov–Bohm phase that electrons acquire in the interferometer, δ is the phase difference of the two alternative paths at zero magnetic field, and $\Delta\theta_{\mathrm{r/t}}$ are the changes in the reflection/transmission phases for the detector electron depending on which path the electron in the interferometer takes. This expression for the transmission can be rewritten in the form

$$\mathcal{T}_\mathrm{i} = |a_1|^2 + |a_2|^2 + 2|a_1||a_2|A\cos(2\pi\phi/\phi_0 + \tilde{\delta}), \tag{21.3}$$

with the decoherence factor A given by

$$A^2 = (|r_1||r_2| + |t_1||t_2|)^2 - 4|r_1||r_2||t_1||t_2|\sin^2\frac{\Delta\theta_\mathrm{r} - \Delta\theta_\mathrm{t}}{2}.$$

The factor A will always obey $0 \le A \le 1$. In order to get more insight into this expression, we write $t_i = \sqrt{\mathcal{T}_i}$, $|r_i| = \sqrt{\mathcal{R}_i} = \sqrt{1-\mathcal{T}_i}$ $(i = 1, 2)$, $\mathcal{T} \equiv \mathcal{T}_1 = \mathcal{T}_2 + \Delta\mathcal{T}$, and $\gamma = \sin[(\Delta\theta_\mathrm{r} - \Delta\theta_\mathrm{t})/2]$ and obtain

$$A^2 = \left(\sqrt{(1-\mathcal{T})(1-\mathcal{T}+\Delta\mathcal{T})} + \sqrt{\mathcal{T}(\mathcal{T}-\Delta\mathcal{T})}\right)^2 - 4\sqrt{(1-\mathcal{T})(1-\mathcal{T}+\Delta\mathcal{T})\mathcal{T}(\mathcal{T}-\Delta\mathcal{T})}\gamma^2.$$

Again we see that the detector can obtain information about the path of the electron in the interferometer in two distinct ways: (1) its transmission depends on which path the electron takes (first term), and (2) its phases $\Delta\theta_\mathrm{r} - \Delta\theta_\mathrm{t}$ depend on which path the electron takes (second term). As an example, we plot in Fig. 21.14 the factor A for the case of $\mathcal{T} = 1/2$, i.e., where the quantum point contact is most sensitive for charge detection.

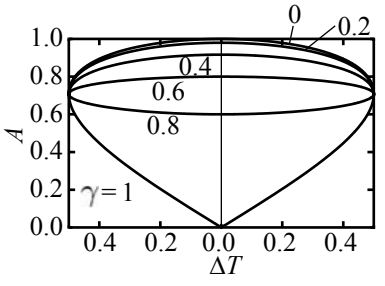


Fig. 21.14 Decoherence factor A as a function of $\Delta\mathcal{T}$ for $\mathcal{T} = 1/2$ and a number of values for γ.

Let us analyze the expression for A by expanding it for small ΔT, i.e., for weak interferometer–detector interaction. It yields

$$A^2 = 1 - 4\mathcal{T}(1-\mathcal{T})\gamma^2 + 2(1-2\mathcal{T})\Delta T\gamma^2 - \frac{(1-2\gamma^2)\Delta\mathcal{T}^2}{4T(1-\mathcal{T})}. \tag{21.4}$$

There is, in lowest order, an interference oscillation amplitude reduction related to the shot noise in the detector [factor $\mathcal{T}(1-\mathcal{T})$] acting back on

the interferometer when retrieving which-path information from phase changes in the detector (via γ). If, by coincidence or design, no such phase changes are present ($\gamma = 0$), then the first dephasing term is of second order in $\Delta\mathcal{T}$ (Buks *et al.*, 1998).

Many detector electrons. So far we have only considered a single electron in the detector quantum point contact. However, it turns out that the relevant quantity is the number of detector electrons interacting with the electron in the interferometer during its passage from the entrance to the exit of the interferometer, i.e., its dwell time τ_i. If we consider electron states in the quantum point contact to be represented by a stream of wave packets, then there is the attempt frequency $2eV/h$ (factor 2 by assuming spin degeneracy) at which electrons in the detector probe the tunneling barrier of the quantum point contact, if V is the voltage applied between its source and drain. Therefore the number of electrons in the detector sensing which-path information of a single electron in the interferometer is $N = 2eV\tau_\mathrm{i}/h$. The decoherence factor A in eq. (21.3) then has to be replaced by A^N (Averin and Sukhorukov, 2005; Neder *et al.*, 2007*a*). If $N \ll 1$, decoherence becomes irrelevant, as the detector is not able to extract significant information about the interferometer.

Decoherence rate. We can introduce the measurement-induced decoherence rate τ_φ^{-1} by equating $\exp(-\tau_\mathrm{i}/\tau_\varphi) = A^N$ and obtain

$$\frac{1}{\tau_\varphi} = -\frac{|e|V}{h}\ln A^2.$$

In the weak coupling limit where $\Delta\mathcal{T}$ is small, we may only consider the first two terms in the expansion (21.4) and obtain the decoherence rate

$$\frac{1}{\tau_\varphi} = -\frac{eV}{h}\ln(1 - 4\mathcal{T}(1-\mathcal{T})\gamma^2)$$

which simplifies for small γ^2 to

$$\frac{1}{\tau_\varphi} = \frac{eV}{h}4\mathcal{T}(1-\mathcal{T})\gamma^2.$$

Using eq. (20.19) for the zero temperature shot noise of a quantum point contact we find the decoherence rate to be directly proportional to the noise power spectral density

$$\frac{\hbar}{\tau_\varphi} = \frac{\gamma^2}{\pi}\frac{h}{e^2}\tilde{S}(\nu),$$

and the parameter $R_\mathrm{c} = h\gamma^2/\pi e^2$ can be interpreted as the coupling impedance between the interferometer and the detector with γ being a dimensionless coupling parameter. Through the above considerations we have now seen how a number of different topics of this book, namely interference, decoherence, and shot noise, are closely related and merge into a consistent picture of decoherence by measurement.

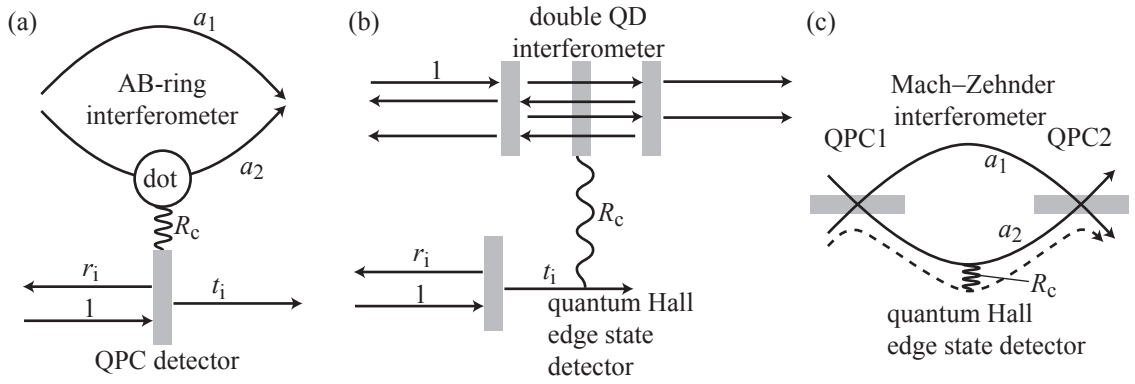


Fig. 21.15 (a) Schematic setup in which an Aharonov–Bohm interferometer is coupled capacitively to a quantum point contact detector. (b) Schematic setup in which a double quantum dot interferometer is coupled capacitively to a partitioned quantum Hall edge state. (c) Schematic setup in which a Mach–Zehnder interferometer is capacitively coupled to a partitioned quantum Hall edge state.

Experiments. Figure 21.15 shows schematically the three mesoscopic arrangements that have been used in experiments to demonstrate controlled decoherence. In the experiment of Buks *et al.*, 1998, shown in (a), an Aharonov–Bohm double slit interferometer with a quantum dot embedded in one arm has been coupled to a quantum point contact detector. The quantum dot has been introduced in one arm in order to increase the dwell time of interferometer electrons in the lower arm and allow for a longer interaction time. During this interaction time many electrons pass the quantum point contact detector thereby enhancing the strength of the decoherence. Nevertheless, the setup stayed in the regime of very weak coupling. The visibility of Aharonov–Bohm oscillations was reduced by 0.3% at most as a result of detector operation.

The most recent experiment reported in Neder *et al.*, 2007*a*, is shown in Fig. 21.15(c). It operates in the quantum Hall regime and uses the electronic Mach–Zehnder interferometer discussed in section 16.5. While the filling factor $\nu = 1$ edge channel is used in the Mach–Zehnder interferometer, a $\nu = 2$ edge channel is partially occupied (partitioned) along one arm of the interferometer and therefore creates shot noise capable of dephasing the interferometer. Coupling between electrons in the two edge channels is via the electrostatic Coulomb interaction. Although this experiment is very interesting we will not further discuss it here, as details of the experiment are still under debate and investigation. We refer the interested reader to the original papers.

Instead we will have a closer look at the experiment reported in Sprinzak *et al.*, 2000, shown schematically in Fig. 21.15(b). In this experiment the interferometer consists of a double quantum dot system. We have seen in section 18.3.1 that a single quantum dot can be regarded as the electronic version of a Fabry–Perot interferometer. The same is true in principle for double quantum dots, but the sum over interfering paths cannot easily be performed as in the single dot case. However, it has

been shown experimentally that double quantum dot resonances are not broadened by the thermal smearing of the Fermi–Dirac occupation function in the leads but rather by their intrinsic decay rate Γ_i (Livermore *et al.*, 1996). This allows us to observe the increase in the resonance width caused by decoherence. The detector in this experiment is a partitioned quantum Hall edge channel. The partitioning is achieved by sending the edge channel through a distant quantum point contact with tunable transmission. There is no direct Coulomb interaction between the double quantum dot and this quantum point contact. The presence of the electron in the interferometer is merely sensed by a phase change of the edge channel state in the vicinity of the double dot. Therefore, this experiment realizes the interesting case $\Delta\mathcal{T} = 0$ with finite γ, because an electron in the edge channel cannot be backscattered. Of course, it is not possible in this setup to read out the detector phase information through a conductance measurement. In order to do that one would have to perform a Mach–Zehnder-like edge channel interference experiment, which is not done here. The decohering effect of the shot noise in the partitioned edge channel is rather measured as an increase of the full-width-at-half-maximum contour, and as the decrease of the peak height of a double quantum dot conductance resonance. More specifically the contour area and the peak height are determined at a triple point in the plane of the two double quantum dot plunger gates.

Figure 21.16(a) shows the conductance of the double quantum dot interferometer as a function of the two plunger gate voltages V_pg1 and V_pg2. We see the familiar charge stability diagram with its hexagon pattern. A single triple point is magnified in (b).

Figures 21.16(c) and (d) show the main results of the experiment. The area A of the full-width-at-half-maximum contour line of the resonance shown in (b), which is taken as a measure of τ_φ^{-1}, exhibits a pronounced maximum around the edge channel partitioning $\mathcal{T}_\mathrm{d} = 1/2$, as expected from the $\mathcal{T}_\mathrm{d}(1-\mathcal{T}_\mathrm{d})$-dependence of the shot noise power spectral density. The inset demonstrates that this area increases roughly linearly with the voltage V_d applied to the partitioning quantum point contact, as expected. In a similar fashion, the height of the triple point conductance peak shown in (d) is minimum, where A is maximum. This behavior is interpreted as convincing evidence for shot-noise-induced decoherence of the double quantum dot interferometer.

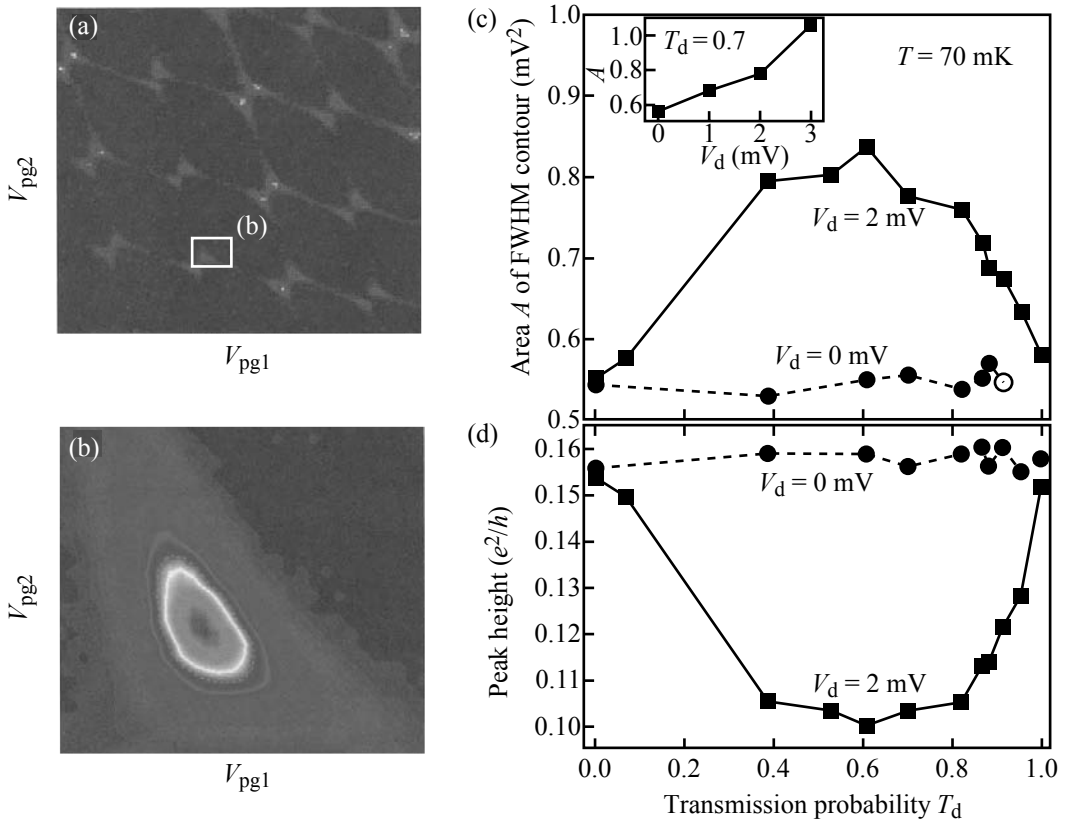


Fig. 21.16 (a) Charge stability diagram of the double quantum dot as a function of the two plunger gate voltages V_{pg1} and V_{pg2} showing the familiar hexagon pattern. The grayscale represents the linear conductance through the double quantum dot. A single triple point is marked with a white rectangle indicating the zoom shown in (b), where the contour line from which the area A is determined can be discerned. (c) Area A of the full-width-at-half-maximum contour as a function of detector partitioning $\mathcal{T}_d$. The inset shows the evolution of the area A as a function of voltage applied to the quantum point contact. (d) Conductance peak height at the triple point as a function of detector partitioning $\mathcal{T}_d$. (Reprinted with permission from Sprinzak *et al.*, 2000. Copyright 2000 by the American Physical Society.)

Further reading

- Papers Fano effect: Fano 1961; Gores *et al.* 2000; Johnson *et al.* 2004; Kobayashi *et al.* 2002.
- Papers transmission phase measurements: Schuster *et al.* 1997.
- Papers controlled dephasing: Buks *et al.* 1998; Sprinzak *et al.* 2000; Neder *et al.* 2007*a*.

Exercises

(21.1) The interference in the interferometer–detector arrangement is reduced by the factor $r_1 r_2^\star + t_1 t_2^\star$. Discuss the physical meaning and the implications of two extreme cases:

(a) $r_1 = r_2^\star$, and $t_1 = t_2^\star$.

(b) $r_1 = 0$, and $r_2 = 1$.

(21.2) Reconsider the general interferometer–detector arrangement depicted in Fig. 21.13. Starting from the entangled state between the two subsystems, take its magnitude squared and integrate out the interferometer variable x. Discuss the meaning and the physical implications of the classical and interference contributions of the result. Think of an experimental arrangement in which the interference contribution can be measured.

Quantum information processing

22

In this chapter we set out to touch a huge field of research which has also been the driving force between a number of beautiful transport experiments on semiconductor nanostructures in recent years. However, the topic is extremely broad, such that we can only give a short overview with selected examples. The chapter is divided into three sections of which only the last section is really related to quantum information and related experiments. The first two sections are a little detour into the field of classical information and its relation to physics that the author felt necessary to include, because most physics curricula do not incorporate the notion of information and its relation to physical systems. A reader who is only interested in the physics of semiconductor nanostructures can proceed directly to the last section of this chapter.

Information processing is a very general term comprising many different situations. One of them is analog information transmission, such as the radio-frequency transmission of the news from the radio station to our homes via electromagnetic waves. Similar is the digital information transmission such as that from my keyboard to the main board of my computer. In the language of information processing, these two situations would be called *communication* (although ordinary people may prefer to call it a one-way data transmission). Another type of analog information processing is the conversion of a current signal measured in an electrical circuit to an output voltage in a current–voltage converter. This type of information processing may be called an analog calculation, as the output voltage is essentially the input current multiplied by the feedback resistor. More complex circuits performing analog calculations may result in analog computing. On the other hand, logic circuits perform logic operations on digital input signals. This would be called a digital calculation, or in a more complex context, digital computation. In many instances, mixtures of analog and digital information processing work together. If I speak into my mobile phone, my voice reaches the microphone as an analog pressure variation, the microphone translates it into a time-dependent voltage, and this electronic signal is then digitized, and so on.

It is important to realize at this point that information processing of any kind is always based on physical systems and physical processes. As time evolves, the physical system evolves according to physics laws and, during this evolution, what we perceive as information is transformed,

and sometimes lost. Information must therefore be recognized as a characteristic quantity related to physical systems. As Landauer phrased it: 'information is physical'.

Because most of our daily experience is (even for us experimental physicists) governed by the laws of classical physics, man-made information processing devices have so far been based on these classical laws, and we therefore talk about classical information processing. As the size of electronic devices used for information processing shrinks down to the mesoscopic- or nano-scale, or perhaps even down to the atomic scale, and the intensity of radiation used for information transmission decreases more and more thanks to more sensitive receivers, it is natural to ask whether the quantum laws governing this world bring about any changes in the way we can process information. Can we still send a radio program by sending a very dilute stream of single photons? Can we still do calculations and computations with individual electrons, or spins? Nature herself has certainly brought about means of information processing that are closer to these ideas, for example, by encoding the genetic information in DNA molecules and finding smart ways of processing this information. Thoughts along these lines have led scientists to work out a novel theory for information processing which has been called *quantum information processing*, and in the meantime experimentalists have started to put some of these concepts into practice.

Because the notion of information is so far missing from the basic education of many (if not most) physicists, we will try to discuss this topic in some detail below, always emphasizing the close relation between information processing and its physical implementation.

22.1 Classical information theory

22.1.1 Uncertainty and information

The notion of classical *information* can be derived by considering probabilistic experiments and their statistical distribution functions (Shannon, 1948). It is directly related to the notion of *uncertainty*. Although the following considerations may seem to be quite abstract and remote from the physical world, we emphasize that Shannon was led to his theory of information by considering physical implementations of the communication of his time.

We consider a probabilistic experiment such as, for example, throwing dice, or drawing a card in a card game. Before the experiment, we do not know its result. There is an *uncertainty* about the outcome. We describe this uncertainty by assigning a probability distribution to the possible outcomes of the experiment. Assume there are Ω possible outcomes that can all occur with the same probability $p = 1/\Omega$. The larger the number Ω of possible outcomes, the bigger is the *uncertainty* U_{before} about the outcome of the experiment. After the experiment we know the result and our *uncertainty* has disappeared. We denote this with $U_{\text{after}} = 0$.

We now define the obtained *information* ΔI as

$$\Delta I = U_{\text{before}} - U_{\text{after}}.$$

Viewed in this way, the information and the uncertainties are real numbers. But how do we define the *uncertainty* before or after the experiment? The uncertainty before the experiment is larger the larger Ω, i.e., the smaller p is. One possibility would therefore be to define the uncertainty to be proportional to $\Omega = 1/p$.

However, such a definition is not in agreement with our intuition about uncertainty. If we conduct, for example, two similar experiments having Ω outcomes each, we would expect the uncertainty to be twice as high as for a single experiment. But the total number of possible outcomes of the double experiment is Ω^2. Therefore it is more intuitive to define the obtained information as the logarithm of the number of possible outcomes, i.e.,

$$U_{\text{before}} = k \ln \Omega = -k \ln p, \tag{22.1}$$

where k is a constant that can be used to define the units of uncertainty and information. Equation (22.1) is also known as the Hartley function, or Hartley entropy (Hartley, 1928). After the experiment $\Omega = 1$, i.e., we know the result. For this case, our definition leads us to $U_{\text{after}} = 0$. In the case of two possible outcomes (e.g., 0 and 1) we call the uncertainty to be '1 bit',[1] i.e.,

$$k \ln 2 = 1 \,\text{bit}.$$

This defines the constant to be $k = 1/\ln 2\,\text{bit}$ and we can write

$$U = \log_2 \Omega \ \text{bit} = -\log_2 p \ \text{bit}.$$

As an alternative, sometimes the unit '1 digit' is used for the quantity information. This corresponds to the case of ten possible outcomes such that each result can be labeled with one of the digits $0 \ldots 9$. Therefore

$$1 \ \text{digit} = k \ln 10 = k \ln 2 \log_2 10 \approx 3.322 \ \text{bit}.$$

Table 22.1 shows the relation between commonly used bit numbers and digits.

The above definition (22.1) of the uncertainty is valid for the special case that all outcomes of our experiment have the same probability. How do we generalize the definition for Ω possible outcomes with possibly different probabilities $p_1, p_2, p_3, \ldots, p_\Omega$? The probabilities obey the sum rule

$$\sum_{n=1}^{\Omega} p_n = 1.$$

In order to find the generalized definition of the uncertainty we consider a simple example: we place $N_0 \gg 1$ cards with the number 0, and

[1] The notion of the 'bit' was introduced by J.W. Tukey and is an abbreviation of 'binary digit'.

Table 22.1 Conversion between commonly used bit and digit numbers.

bits	digits
1	0.30
3.32	1
6.64	2
8	2.41
9.97	3
10	3.01
12	3.61
13.3	4
16	4.82
16.6	5
19.9	6
20	6.02

$N_1 \gg 1$ cards with the number 1 in an arbitrary sequence face down in a row of length $N = N_0 + N_1 \gg 1$. How much information do we gain if we turn the cards in the row and uncover the numbers. The number of possible outcomes corresponds to the number of possibilities to distribute N_0 zeros on N places, i.e.,

$$\Omega = \frac{N!}{N_0!(N-N_0)!} = \frac{N!}{N_0!N_1!}.$$

The information gain, i.e., the uncertainty before learning the sequence is according to our definition

$$U_{\text{before}} = k \ln \Omega = k \left(\ln N! - \ln N_0! - \ln N_1! \right).$$

For sufficiently large N_0 and N_1 we can approximate the right-hand side using Stirling's formula $\ln N! \approx N(\ln N - 1)$ and we obtain after a little algebra

$$U_{\text{before}} = k \ln \Omega \approx -kN \left[\frac{N_0}{N} \ln \frac{N_0}{N} + \frac{N_1}{N} \ln \frac{N_1}{N} \right].$$

If we denote the relative frequencies $p_0 = N_0/N$ and $p_1 = N_1/N$, we obtain the *mean* uncertainty per card

$$\frac{U_{\text{before}}}{N} \approx -k \left[p_0 \ln p_0 + p_1 \ln p_1 \right] = - \left[p_0 \log_2 p_0 + p_1 \log_2 p_1 \right] \text{ bit}.$$

The generalization of this example to cards with more than two different numbers is completely analogous. We assume that each card carries a number between 0 and $n-1$. The number $0 \leq i < n$ occurs N_i times and the number of cards is $N = \sum_{i=0}^{n-1} N_i$. The number of possible sequences is then

$$\Omega = \frac{N!}{\prod_{i=0}^{n-1} N_i!}$$

and the uncertainty is

$$U_{\text{before}} = k \left(\ln N! - \sum_{i=0}^{n-1} \ln N_i! \right).$$

Using Stirling's formula we obtain

$$U_{\text{before}} \approx -kN \sum_{i=0}^{n-1} \frac{N_i}{N} \ln \frac{N_i}{N},$$

and the mean uncertainty per card is

$$\frac{U_{\text{before}}}{N} \approx -k \sum_{i=0}^{n-1} p_i \ln p_i = - \sum_{i=0}^{n-1} p_i \log_2 p_i \text{ bit}.$$

This formula was first used by C.E. Shannon as a definition for the uncertainty (Shannon, 1948), i.e.,

$$H(\{p_i\}) \equiv \frac{U}{N} = - \sum_{i=0}^{n-1} p_i \log_2 p_i \text{ bit}. \tag{22.2}$$

This quantity is called *Shannon entropy*. It is a function of the probability distribution $\{p_i\}$, i.e., one can always calculate it for any given probability distribution. In the case of constant $p_i = 1/n$ and $N = 1$ we again obtain our initial definition (22.1). Equation (22.2) can be interpreted as follows: the quantity $-\log_2 p_i$ is the information gain when symbol i is uncovered. If this symbol arises N_i times within the sequence, we gain the information $-N_i \log_2 p_i$ from these symbols and the total information from the sequence is $-\sum_i N_i \log_2 p_i$. Because for long sequences $N_i/N \approx p_i$, we have for the total information $-N \sum_i p_i \log_2 p_i$, and for the average information per symbol $-\sum_i p_i \log_2 p_i$.

We mentioned at the beginning of this chapter that we often describe our uncertainty by assigning a probability distribution $\{p_i\}$. The Shannon entropy characterizes this set of numbers by distilling a single number out of it. We will see in the examples below that this single number has a certain significance.

Example: Entropy of the binomial distribution. As an application example of the above considerations, we calculate the Shannon entropy of the binomial distribution

$$p_k = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},$$

which has the expectation value $\mu = \langle k \rangle = np$ and the variance $\sigma^2 = \langle (k-\mu)^2 \rangle = np(1-p)$. Inserting the distribution into the definition (22.2) results in

$$H = \log_2(\sqrt{2\pi e}\sigma) + \mathcal{O}(n^{-1}). \tag{22.3}$$

For sufficiently large n, the entropy of the binomial distribution is therefore given by the standard deviation σ. In this limit the binomial distribution becomes the Gauss distribution with the mean value μ and the standard deviation σ. The broader this distribution is, the bigger is the uncertainty about the value of k.

22.1.2 What is a classical bit?

In the above discussion we have introduced the bit as a kind of unit of measurement for the Shannon entropy. One bit of information is obtained as the answer to a yes/no-question. Expressing this answer in numbers, we could use 0 and 1. Binary numbers can be represented by strings of zeros and ones. One position in such a string is called a bit.

Following the discussion of Mermin (Mermin, 2007) the state of a bit can, in analogy to quantum mechanics, be described by two state vectors in Dirac notation, e.g., $|0\rangle$, and $|1\rangle$. Sequences of bits (bit registers) may then be described as tensor products of the state vectors of individual bits. For example, the sequence 0101 would be written as

$$|0\rangle \otimes |1\rangle \otimes |0\rangle \otimes |1\rangle \equiv |0101\rangle .$$

When we manipulate bits during a calculation, the vector notation

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

can be very useful because operations can then be represented by multiplying matrices and vectors. This notation is related to the Pauli notation of two-component wave functions.

If we write a string of zeros and ones on a piece of paper, the bit becomes physical reality. It is common practice to use the term 'bit' also for the physical system that represents zero or one. In general, such a physical bit must have two clearly distinguishable states which can be determined by a read-out measurement. In case of the bit string written with a pencil on a piece of paper, we can perform the read-out measurement with our naked eyes. In more involved implementations of physical bits, the read-out measurement uses some technical apparatus. One nice example is bits implemented on our computer hard discs as regions in which the material is magnetized in a particular direction. The read-out measurement is performed by a read head scanning the surface and detecting the direction of the magnetization by measuring the resistance of a special material showing the so-called giant magnetoresistance.

Classical bits can be read many times without changing their states significantly by the reading process. Information is not lost during the reading process. Classical bits can also be copied or reproduced a large number of times. Of course, degradation of the information may occur in real physical systems, but it occurs after large numbers of reading or copying processes.

The two states of classical physical bits (sometimes called 'Cbits') can be distinguished by measuring a macroscopic state variable, such as magnetization, or a voltage. A vast number of different microscopic states of the bit would give the same value for the macroscopic state variable. The time evolution of classical physical bits is therefore naturally described with the laws of statistical mechanics and thermodynamics. We will therefore have a closer look at the relation between information and thermodynamics a bit later.

In present day computers, information storage, transmission, and processing requires, in an abstract physics language, two things:

(1) A phase space on which statistical ensembles are defined. The state of a particular system at a given time is a point in phase space. This phase space can be discrete, or continuous.

(2) A probability distribution defined on the phase space describing our uncertainty about the state of the system.

Typically physical bits are not in thermodynamic equilibrium with their surrounding. The physical interaction processes decide, for example, about the permanence of the stored data. If the data storage device approaches a state of thermodynamic equilibrium with its surrounding, the stored information is inevitably lost.

Also data transmission is achieved using physical methods and systems. The easiest way to transmit information is by physically transporting a memory device. Often electromagnetic waves are used for information transmission, but in other cases we use massive objects (think about this book that you may have carried from the bookstore to your home), or sound waves (your speech). It is therefore generally believed that information cannot be transmitted faster than the speed of light.

22.1.3 Shannon entropy and data compression

The Shannon entropy has an important meaning beyond quantifying uncertainty. It tells us how many physical bit-resources we need in order to store the stream of data of a data source. As a simple example, we consider a source producing the four symbols 1, 2, 3, and 4. In order to save the information contained in a stream of these four symbols, we could use two bits. We assume here that the source produces the four symbols with different probabilities. For example, let $p_1 = 1/2$, $p_2 = 1/4$, and $p_3 = p_4 = 1/8$. In this case we can reduce the number of bits needed to store a stream of data by using for the frequent symbol 1 fewer bits than for the more rarely occurring symbols 3 and 4. One possibility for encoding the symbols in a chain of bits is given in Table 22.2. Given a very long sequence of the four symbols the mean number of bits per symbol is given by $1/2 \cdot 1 + 1/4 \cdot 2 + 1/8 \cdot 3 + 1/8 \cdot 3 = 7/4$ bits, i.e., less than two. Most interestingly, this result coincides exactly with the Shannon entropy:

$$H(1/2, 1/4, 1/8, 1/8) = \\ -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{8}\log_2\frac{1}{8}\right) = 7/4.$$

It can be shown that the Shannon entropy quantifies how many bits are required *at least* for storing a stream of data. An encoding that uses exactly this number of bits realizes the maximum possible data compression (in the limit of very long data streams). Any further compression would lead to loss of data.

Table 22.2 Encoding of the sequence of numbers $1 \ldots 4$.

1	0
2	10
3	110
4	111

22.1.4 Information processing: loss of information and noise

Logical elements from which machines for information processing are made can be treated like a communication process. Both may have n inputs and m outputs, as shown in Fig. 22.1. As physicists, we can regard any kind of time evolution of a physical system as a communication or information processing channel. In this case, the input signal would be the initial state of the physical system (e.g., a calculator), the output signal would be its final state.

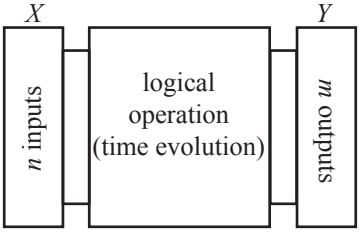


Fig. 22.1 Schematic representation of a communication process, or a logical operation having n inputs represented by the variable X, and m outputs, represented by the variable Y.

Choosing a slightly more abstract example, we may consider an input signal X consisting of five bits. Then, there are $2^5 = 32$ different input

signals (values of X). Each input signal creates with a certain probability an output signal Y. If we take, for example, the output to have 4 bits, then there are $2^4 = 16$ different output signals. The probability, that a particular input signal x turns into the output signal y is described by the conditional probability $p(y|x)$. It describes the probability to find the output signal y, given that the input signal is x. The conditional probability obeys the sum rule

$$\sum_y p(y|x) = 1, \tag{22.4}$$

because each possible input signal creates an output signal. The conditional probabilities $p(y|x)$ can be represented as an $m \times n$-matrix $P(Y|X)$. The probability distributions for the output states can be obtained from

$$p(Y) = P(Y|X)p(X). \tag{22.5}$$

Relating the physical processes (time evolution of the physical system) occurring in information processing to this matrix of conditional probabilities makes the link between physics and information theory.

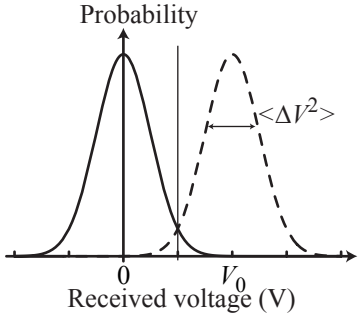


Fig. 22.2 Voltage distribution measured by the receiver if the source is in the state 0 (solid line), or one (dashed line). The thin vertical line discriminates in the interpretation of the measurement between the zero and the one state.

Example I: digital communication channel. A digital communication channel transmits a stream of bits from a data source to a receiver. This can, for example, be implemented by sending light pulses down an optical fiber, or by sending electromagnetic voltage pulses along a coaxial cable. Another example could be data storage: the data source writes a stream of bits into the magnetic layer of a magnetic tape. The receiver reads the data from this medium, perhaps many years, or even decades later. All implementations of this kind will have one input ($n = 2^1 = 2$), and one output ($m = 2$). The matrix of the conditional probabilities is

$$P(Y|X) = \begin{pmatrix} p(0|0) & p(0|1) \\ p(1|0) & p(1|1) \end{pmatrix}.$$

An *ideal* communication channel would have $p(0|0) = p(1|1) = 1$, and $p(0|1) = p(1|0) = 0$, leading to a 2×2 unity matrix, and meaning that the information is perfectly transmitted (or, in case of the magnetic tape, stored without data loss). Such a case, where all elements of $P(Y|X)$ are either zero or one, is called *deterministic*. However, any physical implementation of a communication channel will suffer from some kind of noise. If we choose to represent the two states of a bit with two values of a voltage (say, $0 \equiv 0\,\mathrm{V}$, $1 \equiv V_0$), the voltage source will always be noisy. In the best case, this will only be the thermal noise. Furthermore, the receiver, a voltmeter, will suffer from its own input noise. To be even more specific, assume that the total noise seen by the voltmeter has a gaussian distribution of width $\langle \Delta V^2 \rangle^{1/2}$ (the average voltage noise). The ratio $V_0^2/\langle \Delta V^2 \rangle \equiv S/N$ is called the *signal-to-noise ratio*. Setting the voltage source to 0 V would lead to the distribution of measured voltages shown in Fig. 22.2 as a solid line. Likewise, setting the voltage source to V_0 would lead to the dashed line. In a typical experimental situation, the

voltage noise $\langle \Delta V^2 \rangle$ is proportional to the bandwidth Δf of the system. In order to infer the state of the sender from a single measurement, the receiver needs to use a criterion, which measured voltage will be interpreted as a zero, and which will be interpreted as a one. We may choose the rule that any measured voltage $V \geq V_0/2$ is interpreted as a one, and $V < V_0/2$ is interpreted as a zero. Obviously, there is a chance that the measurement will be interpreted in the wrong way, i.e., that the source sends a 0, but the reciever interprets it as 1, and vice versa. Because in this example the situation is symmetric, we have $\alpha = p(0|1) = p(1|0) \neq 0$. The chance α to make a wrong interpretation can be obtained by integrating the dashed gaussian distribution from $-\infty$ to $V_0/2$. The result depends on the signal-to-noise ratio S/N as shown in Fig. 22.3. The corresponding matrix describing the data transmission would be

$$\begin{pmatrix} p(0|0) & p(0|1) \\ p(1|0) & p(1|1) \end{pmatrix} = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix}.$$

This situation is not deterministic, but it is called *probabilistic*.

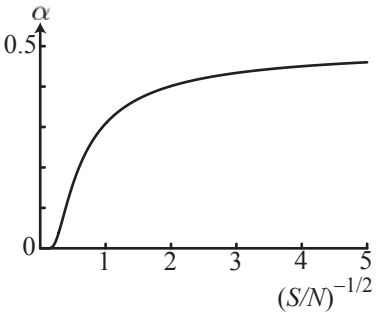


Fig. 22.3 Error probability of a noisy communication channel for the case of symmetric gaussian noise on the zero and one states.

Example II: logical AND gate. The logical AND gate has two inputs ($n = 2^2 = 4$) and one output ($m = 2$). The matrix of the conditional probabilities for an ideal gate without noise is

$$\begin{pmatrix} p[0|(00)] & p[0|(01)] & p[0|(10)] & p[0|(11)] \\ p[1|(00)] & p[1|(01)] & p[1|(10)] & p[1|(11)] \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The gate can be called *deterministic*, because each input state leads to exactly one well-defined output state. A realistic AND gate will, of course, also suffer from noise, and the zeros in the above matrix would be replaced by the appropriate error quantities that may be obtained from an experimental analysis of the physical system. From the ones in the matrix these error quantities will have to be subtracted in order to fulfill eq. (22.4).

Derived entropies. We are now prepared to introduce a number of entropy quantities related to the Shannon entropy defined in eq. (22.2) that are useful for the description of communication and information processing. These are the conditional entropies, the mutual information, and the joint entropy. These quantities are schematically represented in Fig. 22.4. They will be defined below and their meaning will be illustrated in the remainder of this section.

Conditional entropy and picking up information (noise) from the environment. In a realistic (probabilistic) situation the matrix elements of $P(Y|X)$ will not all be exactly zero or one, but rather show small deviations from these values. In such a case, the uncertainty about the output state for a given input state x is in analogy to the Shannon entropy given by

$$U_x = -\sum_y p(y|x) \log_2 p(y|x).$$

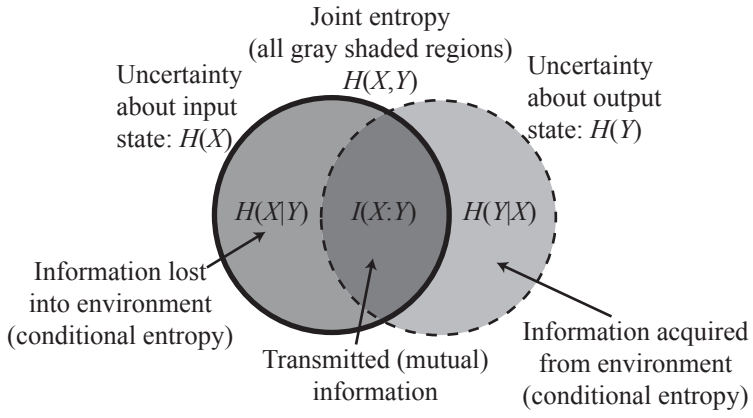


Fig. 22.4 Schematic illustration of the different entropies used in the field of communication and information processing. Physically, communication (or data processing) corresponds to the time evolution of a physical system in 'information space' from the solid circle into the dashed circle. Information from the input may be lost due to the data processing, and additional uncertainty is introduced due to noise in the physical processing apparatus.

If all conditional probabilities $p(y|x)$ were either zero or one (deterministic case) this uncertainty would be zero. If we average this uncertainty over all possible input states x, we obtain a measure of the noise $\mathcal{N}$ of the gate for a given probability distribution of the input states:

$$H(Y|X) = \sum_x p(x)U_x = -\sum_{x,y} p(x,y)\log_2 p(y|x) = \mathcal{N}.$$

Here we have introduced the joint probability $p(x,y) = p(x)p(y|x)$ that the output shows y *and* the input is x. The noise spoiling information processing is described by the conditional entropy $H(Y|X)$, as shown in Fig. 22.4, in analogy to the notion of the conditional probabilities $p(y|x)$ entering it. It describes the information added to the output by the noisy environment, an information we are not interested in when we process the information originating from the source. Alternatively, it may be seen as the additional effort needed to reveal the original information $H(X)$, e.g., by using error detection and error correction schemes. From the definition it follows that $H(Y|X) = \mathcal{N} = 0$ for any deterministic system.

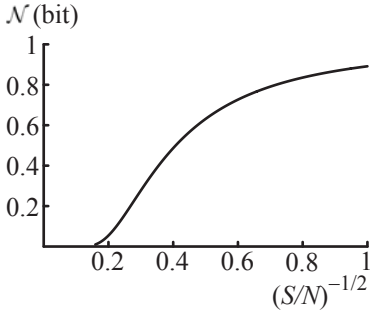


Fig. 22.5 Noise $\mathcal{N}$ of the communication channel with symmetric noise in bits per sent bit as a function of the signal-to-noise ratio S/N.

Example I: probabilistic communication channel. We can apply the concept of the conditional entropy quantifying the noise to our previous paradigmatic example of a communication channel with symmetric noise. In this case, the conditional entropy will depend on the signal-to-noise ratio S/N of the situation with the gaussian distributions shown in Fig. 22.2. The amount $\mathcal{N}$ of useless information about the environment added due to the noise per bit sent is shown in Fig. 22.5, where we have assumed that the source generates the states zero and one with equal probability 1/2. When the signal-to-noise ratio is large, essentially no such information is acquired. When the signal-to-noise level approaches 0, i.e., the separation between the two levels, the environmental noise essentially dominates, and the state of the transmitted bit cannot be identified.

Example II: logical AND gate with noise. As a second example, we consider the model of an AND gate with noise. Consider the conditional probability matrix

$$\begin{pmatrix} p[0|(00)] & p[0|(01)] & p[0|(10)] & p[0|(11)] \\ p[1|(00)] & p[1|(01)] & p[1|(10)] & p[1|(11)] \end{pmatrix} = \begin{pmatrix} 1-\epsilon & 1 & 1 & 0 \\ \epsilon & 0 & 0 & 1 \end{pmatrix}.$$

We assume that all input states occur with the same probability $p(x) = 0.25$. The noise is then given by

$$H(Y|X) = \mathcal{N} = -0.25(1-\epsilon)\log_2(1-\epsilon) - 0.25\epsilon\log_2\epsilon.$$

The value of this function is plotted in Fig. 22.6 as a function of ϵ. The noise has a maximum value of 1/4 bit, when $\epsilon = 1/2$. For $\epsilon \to 1$ the noise reduces, because our possibility to infer an output state for a given input state increases again, but the functionality of the AND gate is not maintained. For $\epsilon = 1$, the input $(0, 0)$ will produce the output 1 with certainty.

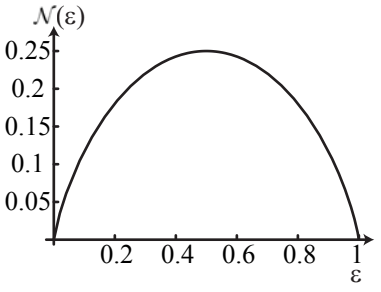


Fig. 22.6 Noise $\mathcal{N}(\epsilon)$ of the AND gate with noise parameter ϵ.

Conditional entropy and losing information into the environment. Beyond noise in a physical system, there are ways of losing (or discarding) information in information processing. We can unravel this, if we investigate the inverse question, namely, what uncertainty we have about the state at the input, if a particular output state y is found. In analogy to the Shannon entropy this is given by

$$U_y = -\sum_x p(x|y)\log_2 p(x|y).$$

The information loss L is now defined as the uncertainty about the input state averaged over all output states with their probability distribution, i.e., as

$$H(X|Y) = L = -\sum_y p(y)\sum_x p(x|y)\log_2 p(x|y) = -\sum_{x,y} p(x,y)\log_2 p(x|y).$$

This is again a conditional entropy (see Fig. 22.4). The probabilities for the output states are given by eq. (22.5). Furthermore, the conditional probabilities obey *Bayes' theorem*

$$p(x,y) = p(y)p(x|y) = p(x)p(y|x). \tag{22.6}$$

Using these two relations we can express the information loss as a function of the probability distribution $p(x)$ of the input states and the conditional probabilities $p(y|x)$ describing the gate:

$$L = -\sum_{x,y} p(x)p(y|x)\log_2 \frac{p(x)p(y|x)}{\sum_\xi p(\xi)p(y|\xi)},$$

where $\sum_\xi$ sums over input states.

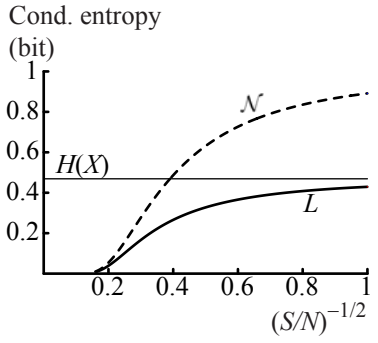


Fig. 22.7 Noise $\mathcal{N}$ and information loss L of the noisy communication channel assuming that the source produces zeros with a probability of 0.9 and ones with a probability of 0.1. The entropy of this input probability distribution (entropy $H(X)$ of the source) is shown as a thin horizontal line.

Example I: probabilistic communication channel. If we assume that the source produces zeros and ones with equal probability 1/2, then the information loss L is exactly equal to the noise $\mathcal{N}$. However, if the source does not produce zeros and ones with equal probability, then $L \neq \mathcal{N}$, as shown in Fig. 22.7. This allows the following interpretation. The loss of information L cannot become larger than the entropy $H(X)$ of the source. The transmitted information (to be exactly defined below) is the difference between $H(X)$ and L (see Fig. 22.4). Added to this transmitted information is the noise which represents information about the physical environment creating it. This part of the entropy of the received distribution $H(Y)$ is useless in terms of information processing. However, the information L lost into the environment cannot be bigger than the entropy of the source.

Example II: logical AND gate (without noise). The ideal noiseless AND gate leaves us with a quite large uncertainty about the input state, if the output reads $y = 0$, because there are three possible input states leading to this result. However, if the output reads $y = 1$, we know with certainty that the input must be $x = (11)$. Assume the four input states of a deterministic AND gate all occur with the same probability $p(x) = 0.25$. Then, the probabilities of the two output states are $p(0) = 0.75$ and $p(1) = 0.25$. With these we calculate for the information loss $L = 0.75 \log_2 3 \approx 1.189$ bit. This information loss is a consequence of the fact that the input consists of two bits, whereas at the output we have reduced the maximum possible information content to 1 bit. Some of the information is therefore dissipated into the environment of the physical system during the operation. This process again complements the addition of information to the output from the environment quantified by $\mathcal{N}$.

Mutual information, transmitted information. Before we know the output state, we are uncertain about the input state according to the Shannon entropy

$$U_{\text{in}} = H(X) = -\sum_x p(x) \log_2 p(x).$$

The information gain about the input state that we obtain from reading the output state can be called the transmitted information M given by

$$I(X:Y) = M = H(X) - H(X|Y) = \sum_{x,y} p(x)p(y|x) \log_2 \frac{p(y|x)}{\sum_\xi p(\xi)p(y|\xi)}.$$

This quantity, also called mutual information $I(X:Y)$, is shown in Fig. 22.4. Following the previous discussion, it is the part of the information at the output that is useful for information processing. The mutual information measures the *correlation* between the input and the output states. It is a more general measure of correlations than Pearson's linear correlation coefficient given by the normalized covariance which is frequently used in data analysis.

If we denote the uncertainty about the output with

$$U_{\text{out}} = -\sum_y p(y)\log_2 p(y),$$

we can show by using the above definitions that

$$I(X:Y) = M = U_{\text{in}} - L = U_{\text{out}} - \mathcal{N} = I(Y:X).$$

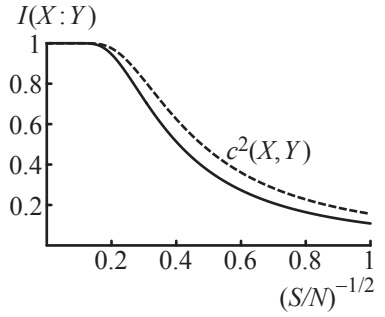


Fig. 22.8 Solid line: Mutual information of a noisy communication channel assuming that the source produces zeros and ones with equal probabilities of 1/2. The dashed line gives the square of Pearson's linear correlation coefficient $c(X,Y)$ for the same situation for comparison.

Example I: probabilistic communication channel. The mutual information transmitted in our example of the noisy communication channel is shown in Fig. 22.8. It is extremely close to one bit if the signal-to-noise ratio (S/N) is sufficiently large, meaning that the receiver can discern the two states easily. When the signal-to-noise ratio becomes comparable to one or even bigger, the transmitted information goes to zero and most of the information in $H(Y)$ originates from the noisy environment. The figure also shows the square of the linear correlation coefficient for comparison. Although its qualitative behavior is (at least in this example) similar to that of the mutual information, it differs quantitatively showing that it is a measure for correlation different from the mutual information. This example teaches us that to communicate means to establish correlations between a sender and a receiver.

Example II: logical AND gate (without noise). In the above case of the ideal AND gate, we have $U_{\text{in}} = 2\,\text{bit}$ (if all $p(x) = 0.25$), $L \approx 1.189\,\text{bit}$, and therefore $I(X:Y) = M \approx 0.811\,\text{bit}$. Here, the mutual information is below one bit because we dump some of the input information in the physical environment. The value is even below one bit, because the probability distribution $p(y)$ deviates from an even distribution.

Joint entropy. In accordance with the intuitive Fig. 22.4 we define the joint entropy

$$H(X,Y) = \sum_{x,y} p(x,y)\log_2 p(x,y).$$

It gives the information that we gain on the average if we learn at the same time that the input state takes the value x *and* the output state takes the value y. It can be represented as (cf., Fig.22.4)

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) = H(Y) + H(X|Y) \\ &= H(X|Y) + H(Y|X) + I(X:Y) \\ &= H(X) + H(Y) - I(X:Y). \end{aligned} \quad (22.7)$$

These relations for the entropies are analogous to Bayes theorem (22.6) for probabilities. The joint entropy can be seen as the sum of all information involved in a certain information processing step, comprising the information lost into the environment, the (useless) information acquired from the environment, and the transmitted information. It can

be bigger than the entropy $H(X)$ of the source, and bigger than the entropy $H(Y)$ of the received signal.

A special situation arises when X and Y are mutually independent. This is a situation which is not desirable in communication, because it implies that there is no information transmitted. In other contexts, however, this may be a reasonable assumption, for example, if we consider two (classical) physical systems which are completely uncorrelated. In such cases, the joint probability is given by

$$p(x, y) = p(x)p(y).$$

Comparing to Bayes' theorem (22.6), statistical independence implies that $p(y|x) = p(y)$, and $p(x|y) = p(x)$. It also implies that the joint entropy is the sum of the two subsystems' entropies,

$$H(X, Y) = H(X) + H(Y).$$

Comparing with eq. (22.7) statistical independence also means that $I(X : Y) = 0$, expressing the fact that there are no correlations between the two systems.

Relative entropy. We now make a brief detour into the field of hypothesis testing in order to define the relative entropy. Suppose you have a coin and you wish to decide whether the coin is fair or not. The hypothesis H_0, i.e., fair implies that heads and tails have equal probabilities $1/2$. In order to derive a quantitative measure on which your decision between the hypotheses can be based, we have to specify the hypothesis H_1 that the coin is unfair, by giving values for the probabilities of heads and tails. Assume we choose them to be $2/5$ and $3/5$, respectively. However, the exact values do not play an important role here (we could also have chosen, say, $7/15$ and $8/15$. We will decide between the two hypotheses by tossing the coin n times, obtaining the data set D. It is now reasonable to base our decision on the ratio of the conditional probability that H_0 is true, given D, and the conditional probability that H_1 is true, given the same dataset D, i.e., on $p(H_0|D)/p(H_1|D)$ called the *odds ratio*. If its value is larger than one, we would prefer H_0 over H_1, if it is smaller than one, we prefer H_1 over H_0. The choice of the threshold 1 for our decision seems to be the natural choice, but in principle, we could make an arbitrary choice. According to Bayes' theorem (22.6) we can write for this ratio

$$\frac{p(H_0|D)}{p(H_1|D)} = \frac{p(H_0)}{p(H_1)} \frac{p(D|H_0)}{p(D|H_1)}.$$

The probabilities $p(H_0)$ and $p(H_1)$ describe our knowledge about the fairness of the coin before the experiment. Because we are unbiased, we have no preference for either of the two, so we set $p(H_0) = p(H_1)$. The ratio $p(D|H_0)/p(D|H_1)$ is called the *likelihood ratio* because it describes the ratio of the likelihoods that we obtain the data set D provided that one of the two hypotheses is true. It is now important to notice that

the dataset D consists of a sequence of *independent* tosses, such that the odds ratio can be written as

$$\frac{p(D|H_0)}{p(D|H_1)} = \prod_{k=1}^{n} \frac{p(D_k|H_0)}{p(D_k|H_1)},$$

where each D_k can be either head (h) or tail (t). It is now convenient to base our decision on the logarithm of the odds ratio

$$\log_2 \frac{p(H_0|D)}{p(H_1|D)} = \sum_{k=1}^{n} \log_2 \frac{p(D_k|H_0)}{p(D_k|H_1)},$$

because in this form the contributions of different tosses on the right add up. We would now prefer H_0 over H_1 if this logarithm gives positive values, whereas we prefer H_1 over H_0 if it yields negative values. If our dataset contains n_h heads and n_t tails, we can rewrite this sum as

$$\log_2 \frac{p(H_0|D)}{p(H_1|D)} = \sum_{i=h,t} \log_2 \frac{p^{n_i}(t_i|H_0)}{p^{n_i}(t_i|H_1)} = n \sum_{i=h,t} \frac{n_i}{n} \log_2 \frac{p(t_i|H_0)}{p(t_i|H_1)},$$

where t_i is either (h) or (t), and $p(t_i|H_j)$ is the likelihood that a single toss gives the result t_i, provided that the hypothesis H_j is true.

For large numbers n, the probability $p(t_i|H_0)$ will be the best estimate for the relative frequency n_i/n, if we *assume* that H_0 is correct. We therefore obtain as the best estimate for the logarithm of the odds ratio

$$\log_2 \frac{p(H_0|D)}{p(H_1|D)} = n \sum_{i=h,t} p(t_i|H_0) \log_2 \frac{p(t_i|H_0)}{p(t_i|H_1)}.$$

The probability that we accept H_1 based on the dataset D, although H_0 is correct, is now

$$p(\text{wrong decision}) \propto 2^{-nH(H_0||H_1)},$$

where

$$H(H_0||H_1) = \sum_{i=h,t} p(t_i|H_0) \log_2 \frac{p(t_i|H_0)}{p(t_i|H_1)}.$$

This quantity is called the relative entropy between the probability distributions $p(t_i|H_0)$ and $p(t_i|H_1)$. It tells us how easily we can distinguish the two distributions after n tosses by giving the rate at which the probability of making a wrong decision decays. If the relative entropy is large, we can correctly identify the right hypothesis after a relatively small number of tosses. If it is small, it is hard to distinguish them, and we have to toss a large number of times. If we had chosen the two probability density distributions to be exactly the same, then the relative entropy would assume the value zero.

Relative entropy and mutual information. We obtain more insight into the meaning of the mutual information, if we consider its relation to the relative entropy. For this purpose we first remember that from any joint probability distribution $p(x, y)$ we can obtain the probabilities $p(x)$ and $p(y)$ by *marginalization*, i.e., by

$$p(x) = \sum_y p(x, y) \quad \text{and} \quad p(y) = \sum_x p(x, y).$$

We can therefore write the mutual information as

$$\begin{aligned} I(X:Y) &= H(X) + H(Y) - H(X,Y) \\ &= -\sum_x p(x)\log_2 p(x) - \sum_y p(y)\log_2 p(y) + \sum_{x,y} p(x,y)\log_2 p(x,y) \\ &= -\sum_{x,y} p(x,y)\log_2 p(x) - \sum_{x,y} p(x,y)\log_2 p(y) + \sum_{x,y} p(x,y)\log_2 p(x,y) \\ &= \sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}. \end{aligned}$$

The last expression is the relative entropy telling us how well we can distinguish the probability distribution $p(x, y)$ from $p(x)p(y)$ which assumes X and Y to be statistically independent. This again supports the idea that the mutual information is a measure of the mutual dependence, or the correlation of two systems.

22.1.5 Sampling theorem

In analog communication systems, the transmission of information is limited by the bandwidth. The *sampling theorem* (again due to Shannon) is an important ingredient that we have to use if we want to find out how much information can be transmitted through a noisy communication channel. Consider a time-dependent signal $V(t)$ (for example, a time-dependent voltage) that we transmit from a sender to a receiver. The transmission channel (for example, the coaxial cable connecting sender and receiver) will have a certain bandwidth f_0 making sure that the transmitted signal does not contain any frequencies $f > f_0$. Now here is the sampling theorem:

If a signal does not contain any frequencies $f > f_0$, then the signal is uniquely determined by its values at discrete points in time if the time separation of these points is smaller than $\Delta t = 1/2f_0$.

This is a remarkable statement, as it tells us that a *continuous* signal can be *completely* described by a series of *discrete* values as shown in Fig. 22.9. In order to get some insight into the proof of this theorem we look at the Fourier transform of the signal $V(t)$ which is given by

$$V(f) = \int_{-\infty}^{\infty} dt\, V(t) e^{-2\pi i f t}.$$

The assumption about the bandwidth limitation of the signal implies that $V(f) = 0$ for $|f| > f_0$. If we sample the signal at time intervals

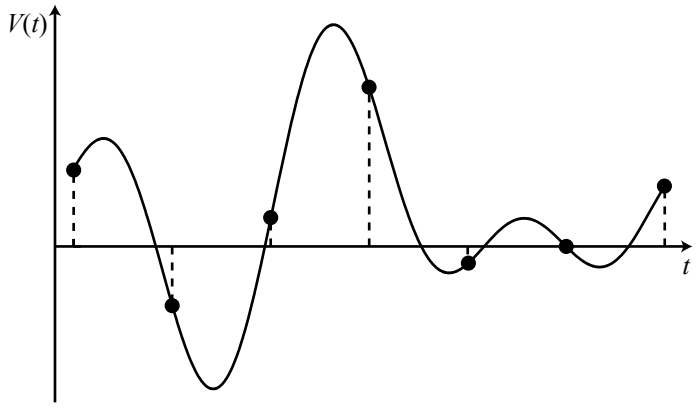


Fig. 22.9 Visualization of Shannon's sampling theorem. The bandwidth limited continuous signal $V(t)$ can be exactly recovered from the precise knowledge of a number of equally spaced discrete data points.

$\Delta t = 1/2f_0$, we obtain the discrete data points $v_n = V(n\Delta t)$. We now introduce the sampled function

$$V_{\mathrm{s}}(t) = \sum_{n=-\infty}^{\infty} V(t)\delta(t - n\Delta t)\Delta t = \sum_{k=-\infty}^{\infty} V(t)e^{2\pi i k f_{\mathrm{s}} t},$$

where $f_{\mathrm{s}} = 2f_0 = 1/\Delta t$ is called the sampling frequency or sampling rate. The sampled function $V_{\mathrm{s}}(t)$ is completely known to the receiver, because it acquires all the v_n. It is now straightforward to show that the Fourier transform $V_{\mathrm{s}}(f)$ of the sampled function is related to the Fourier transform $V(f)$ by

$$V_{\mathrm{s}}(f) = \sum_{k=-\infty}^{\infty} V(f - k f_{\mathrm{s}}).$$

However, due to the limitation of $V(f)$ in bandwidth, terms for different k in the above sum do not overlap. The Fourier transform of the sampled function is therefore simply a repeated copy of the $V(f)$ shifted by integer multiples of f_{s} in frequency. This means that the receiver can recover $V(f)$ from the knowledge of $V_{\mathrm{s}}(f)$ using

$$V(f) = V_{\mathrm{s}}(f)H(f),$$

where $H(f) = 1$ for $|f| < f_0$, and $H(f) = 0$ elsewhere. Reconstruction of the time-dependent signal requires that we transform this equation back into the time domain. This leads to a convolution of the inverse Fourier transforms of $V_{\mathrm{s}}(f)$ and $H(f)$. The latter is given by

$$H(t) = \frac{\sin(\pi t/\Delta t)}{\pi t},$$

leading to

$$V(t) = \int_{-\infty}^{\infty} dt' V_{\mathrm{s}}(t')H(t - t') = \sum_{n=-\infty}^{\infty} v_n \frac{\sin[\pi(t - n\Delta t)/\Delta t]}{\pi(t - n\Delta t)/\Delta t}.$$

In principle, the receiver has reconstructed the signal in the time domain completely, after performing this summation.

A few words of caution are due here concerning the practical implementation of the sampling theorem. Real communication channels may show a bandwidth limitation similar to a low-pass filter. The frequency response of such a filter is, although very small compared to one, not exactly zero beyond the filter bandwidth. The above reconstruction of the signal is therefore better the larger the sampling rate. In practical cases, f_0 is therefore taken to be, for example, 30% larger than the bandwidth of the low-pass. Another practical limitation of the reconstruction procedure is that it requires, in principle, the summation over an infinite number of sampled values, whereas in practice, we sample signals over a limited time interval and therefore obtain only a finite number of samples (see, e.g., Pawlak, 1994). This problem is, however, not so severe because $H(t)$ decays quickly in time for $|t| \to \infty$. Moreover, a continuous signal defined on a limited time interval does, in principle, not have a bandwidth-limited Fourier spectrum. This practical problem is again not critical because the high-frequency Fourier components of a time-limited signal will be very small. Therefore we can state that the sampling theorem is mathematically exact, but its practical implementation is always an approximation of the ideal mathematical setting. Further insights into this topic can be obtained from Slepian, 1976. Nevertheless, the theorem has proven to be of great practical importance in signal processing and image processing. In physicist's labs the theorem is of great use whenever a measurement is performed producing samples of a continuous measurement signal.

22.1.6 Capacitance of a noisy communication channel

In the case of a probabilistic (noisy) communication channel, which we have intensely studied above, the mutual information $I(X : Y)$ tells us how much useful information is received, if 1 bit of information is sent. If we want the receiver to receive a message of n bits, we have to introduce some redundancy in the message. This means that we send a certain number of m bits in order to make sure that $mI(X : Y) = n$. The rate f_s at which we can sample the bits is, according to the sampling theorem, given by the bandwidth f_0 of the system, i.e., by

$$f_\mathrm{s} = 2f_0.$$

Sampling with a higher rate than f_s gives us more information than we need to reconstruct the incoming signal, and sampling with a smaller rate means that we lose information about the input signal. On the other hand, the bandwidth f_0 determines the signal-to-noise ratio in the transmission. Usually we have

$$\frac{S}{N} = \frac{V_0^2}{\langle \Delta V^2 \rangle} = \frac{V_0^2}{S_0 f_0},$$

where S_0 is the spectral density of the voltage noise assumed to be constant here (white noise). The larger the bandwidth f_0, the smaller is the signal-to-noise ratio. However, we have also seen that the signal-to-noise ratio enters the mutual information, i.e, we have

$$I(X:Y) = I\,(S/N) = I\left(\frac{V_0^2}{S_0 f_0}\right).$$

The rate r_i at which the receiver obtains useful information from the sender is therefore given by

$$r_\mathrm{i} = 2 f_0 I\,(X:Y)\,.$$

This is one form of Shannon's noisy channel capacitance theorem. It implies that if a message is sent at a rate r_s it will be corrupted, and error correcting schemes have to be applied. Such schemes require us to introduce redundancy into the information sent, which will effectively reduce the rate of the sent information. If the rate is lowered below r_i, then error correction schemes can be applied in order to reduce the error arbitrarily close to zero. The theorem does, however, not tell us what such an error correction scheme will look like.

Error correction. The simplest error correction scheme would be the repeated transmission of each bit. If, for example, each bit is sent three times, the receiver could decide for the majority, i.e., 110 would be interpreted as 1, whereas 010 as 0.

A different version of the noisy channel capacitance. We arrive at a slightly different version of the above noisy channel capacitance, if we consider the measurement process of a continuous function of time, such as a time-varying voltage $V(t)$. Imagine, for example, that such a signal is created during your measurement on an unknown sample in the lab. Before the measurement you may have an idea about the (gaussian) noise level $\langle \Delta V^2 \rangle$ that your measurement setup produces at its bandwidth f_0. In fact, $\langle \Delta V^2 \rangle = S_0 f_0$, where S_0 is the power spectral density of the noise. In addition, you will perhaps have a rough idea about the voltages you will expect in your measurement, allowing you to identify a gaussian distribution of width V_0 describing the range of expected voltage values. Because the signal and the noise can be expected to be uncorrelated, the two gaussian distributions will simply multiply and give you again a gaussian distribution of width $\sqrt{V_0^2 + \langle \Delta V^2 \rangle}$ describing the expected variation of your input signal. This knowledge allows you not only to choose a proper voltage range on your voltmeter, but also quantifies your uncertainty about the measurement result. According to eq. (22.3) the uncertainty about the result of a single measurement value is given by

$$H = \frac{1}{2}\log_2\left[2\pi e(V_0^2 + \langle \Delta V^2 \rangle)\right].$$

According to the same formula, the information loss resulting from the noise is given by

$$L = \frac{1}{2}\log_2(2\pi e \langle \Delta V^2 \rangle).$$

The information gained about the sample of interest is the transmitted, or mutual information

$$I = H - L = \log_2 \sqrt{\frac{V_0^2 + \langle \Delta V^2 \rangle}{\langle \Delta V^2 \rangle}}.$$

This number answers the question as to how many bits of information we gain about the sample when we take a single measurement point. If the voltage to be measured varies in time, the rate r_i of acquired information is given by the information per data point times the number of data points acquired within a second. The latter is given by the sampling theorem to be $2f_0$, and we obtain

$$r_\mathrm{i} = 2f_0 \log_2 \sqrt{1 + \frac{S}{N}}. \tag{22.8}$$

The rate r_i of bits per second that we acquire during a measurement is fully determined by the bandwidth f_0 of our measurement apparatus, and the signal-to-noise ratio S/N. This is why these two quantities are of such crucial importance in experimental physics labs. Error correction is usually achieved by measuring each data point of a curve many times and averaging the results (another way of looking at this is interpreting the averaging process as a reduction of the bandwidth and the noise). Equation (22.8) is the celebrated version of the noisy channel capacitance theorem that Shannon derived in his theory of communication.

22.2 Thermodynamics and information

We would now like to return to the question of how these theoretical concepts of information are related to the physical world and physical processes used for information processing and storage.

22.2.1 Information entropy and physical entropy

Mathematical similarities. The formulation of the theory of information processing is based on probability theory. The same applies to the notion of entropy used in physics. Both concepts are inherently statistical in nature and applicable only to statistical ensembles, but not to the individual members of the ensembles. In 1872, long before the notion of information became mathematically defined, Ludwig Boltzmann recognized that the problems of the kinetic theory of gases are at the same time problems of probability theory. Boltzmann's H-function

$$H = \sum_i p_i \ln p_i,$$

where the p_i denotes probabilities that particles of a gas are found in a particular state i, has the same form as the negative Shannon entropy (22.2). At the same time Boltzmann's H-theorem

$$\frac{dH}{dt} \leq 0$$

is closely linked to the second law of thermodynamics according to which the physical entropy of an isolated system will always increase if the system starts out of equilibrium and evolves in time. In thermodynamic equilibrium, the entropy reaches its maximum. These statistical ideas were further developed by J.W. Gibbs, M. Planck, J. von Neumann, and others, in the early 20th century.

The modern theory of statistical mechanics describes the entropy of a physical system as

$$S = -k_{\mathrm{B}} \sum_i p_i \ln p_i,$$

where the p_i denotes the probabilities that the system is at equilibrium found in the microscopic state i. If all the microscopic states have the same probability $p = 1/\Omega$, as assumed for the microcanonical ensemble, the entropy reduces to [cf., eq. (22.1)]

$$S = k_{\mathrm{B}} \ln \Omega,$$

where Ω is the number of available microscopic states. The physical entropy can be interpreted as a quantitative measure of our uncertainty about the microscopic state of the system given that the macroscopic state is known. The macroscopic state is described by macroscopic state variables, such as the total energy of the system, the volume that it occupies, and the number of particles in the system. The entropy is a special quantity in statistical mechanics, because it cannot be interpreted as the average of a microscopic property of the system. In contrast, for example, temperature can be defined as the average kinetic energy of the individual molecules in an ideal gas. Entropy is an extrinsic quantity, like the volume and the number of particles.

The definitions of Shannon's entropy and the physical entropy differ in two details: the definition of the physical entropy uses natural logarithms rather than logarithms to the base two, and there is Boltzmann's constant k_{B} that defines the units of the physical entropy to be J/K, in agreement with the historically earlier classical theory of thermodynamics by Clausius. In fact, both differences could be eliminated if we chose to measure temperature in units of energy. This would be, for example, consistent with the kinetic theory of ideal gases, where the temperature is a measure of the average kinetic energy per gas particle.

Relation between physical entropy and information entropy. We have seen above that there are strong mathematical similarities between the concept of information entropy, as introduced by Shannon, and the physical entropy in statistical mechanics. The question, whether there is a deeper relationship between the two, or even whether both describe the same thing, has created a vast amount of research activities over more than 100 years, and a broad agreement on the answer has not emerged yet. A detailed discussion of this topic is certainly much beyond the scope of this book. We will therefore only try to convey briefly a flavor of the many interesting aspects involved.

The first operational use of physical entropy was made by Rudolf Clausius in 1867 (Clausius, 1867), who defined the quantity mathematically, and stated the second law of thermodynamics. In 1871, James Clerk Maxwell published his book 'Theory of Heat', where he discussed the implications of the second law, and suggested a thought experiment in which a 'being with sharpened faculties' would be able to observe individual molecules in a thermally isolated bath which is split into two subsystems A and B that are initially at thermodynamic equilibrium. By observing the molecules and intelligently opening a shutter between the two subsystems, the being can allow the faster molecules from A into B, and the slower ones from B into A thereby heating B as compared to A 'without expenditure of work', and in contradiction to the second law. The whole process indeed *lowers* the total entropy of the joint system, and work could be extracted when the two subsystems are allowed to equilibrate. This thought experiment later became known as Maxwell's demon problem. Since then, researchers have tried to 'exorcize' the demon and to save the second law of thermodynamics. For example, it became clear that a proper treatment has to specify, how the demon *measures* the motion of the molecules and whether it has to invest energy in order to do that, and how this would affect the entropy balance. Some workers have tried to ask which role the intelligence of the demon plays for the entropy balance. In his seminal paper Leo Szilard (Szilard, 1929) made a connection between the *information acquisition* by the demon and the decrease of entropy in the system by suggesting what we now call 'Szilard's engine'. Later work focused on the fact that the demon may need a binary memory for storing the information about the molecules. Adopting the idea that physical entropy and information entropy might be one and the same thing would again save the second law. At some point the demon will need to erase the gathered information from his memory. Landauer showed in 1961 (Landauer, 1961) that the erasure of a bit of information will at least need the energy $k_{\mathrm{B}}T\ln 2$, and at the same time increase the physical entropy of the environment by $k_{\mathrm{B}}\ln 2$. Similar ideas were independently put forward by Oliver Penrose in his 1970 book *Foundations of Statistical Mechanics*, and Landauer's principle, as we call it today, has been proven to be correct in the classical context by a number of other workers more recently, but work is still in progress investigating its validity under extreme quantum conditions. Slightly further on, we will discuss Landauer's principle in more detail.

Information storage in thermodynamic systems. Machines used for classical information processing, such as transistors, operational amplifiers, digital circuits, computers, or even our brain, are thermodynamic systems. The same applies to data storage devices, such as stones on which ancient cultures carved their laws, magnetic tapes, computer hard drives, compact discs, and again our brains.

We give a very simple example which shows how information can be stored in a thermodynamic system. When we have tossed a fair coin with the same probability 1/2 for head or tail, before we look at the

outcome we have an uncertainty about the microscopic state of the coin $U_{\text{before}} = \log_2 \Omega \,\text{Bit} \sim 10^{23}\,\text{Bit}$. After looking at the result, i.e., after measuring whether head or tail is on top, the number of microscopic states compatible with our acquired information is $\Omega/2$. Our uncertainty about the microscopic state of the coin is now $U_{\text{after}} = \log_2(\Omega/2)\,\text{Bit}$. By measuring, we have therefore gained the information

$$\Delta I = \log_2 \Omega \,\text{Bit} - \log_2 \frac{\Omega}{2}\,\text{Bit} = \log_2 2\,\text{Bit} = 1\,\text{Bit}.$$

Alternatively we can say that the information of 1 bit is stored in the orientation of the coin.

A different example for storing information in a thermodynamic system by entropy reduction was given in Feynman, 1996. Information is stored by compressing gas contained in a box of volume V isothermally to the volume $V/2$. Compression into one half of the container corresponds to the state 0, into the other half to 1. The stored information is

$$\Delta I = k \ln \frac{V}{V/2} = 1\,\text{Bit}$$

and the physical entropy is reduced by

$$\Delta S = -k_{\text{B}} \ln 2.$$

The second law of thermodynamics predicts that the physical entropy will increase in time. The gas, for example, will diffuse relatively quickly once the piston used for compression is removed, in order to fill the volume V evenly. The stored information has been lost, and the physical entropy of the system has increased. The time scale over which this happens has to be calculated with microscopic models; it is not given by the second law. If we save information on a magnetic tape using magnetic domains magnetized in a particular direction, the information remains almost unaltered for decades. These large time scales are crucial for storing information with high reliability.

Conceptional parallels between communication and the measurement process. When we discussed the notion of information entropy and its derivatives before, we used a (noisy) communication process as an illustrative example. It has been pointed out that the logical structure of communication is the same as that for making measurements in physics (see e.g., Rothstein, 1951). This analogy is illustrated in Fig. 22.10. The physical system of interest for a measurement corresponds to the information source of communication theory. In the latter case, this could be somebody feeling the urgent wish to tell a remote person something very important. The physicist performing a measurement on the system decides on a particular measurable property, and designs an appropriate measuring apparatus that converts the physical quantity to be measured into another quantity that can be easily transmitted. In communication systems, a corresponding apparatus is used in which any message originating from the source can be encoded and transmitted into

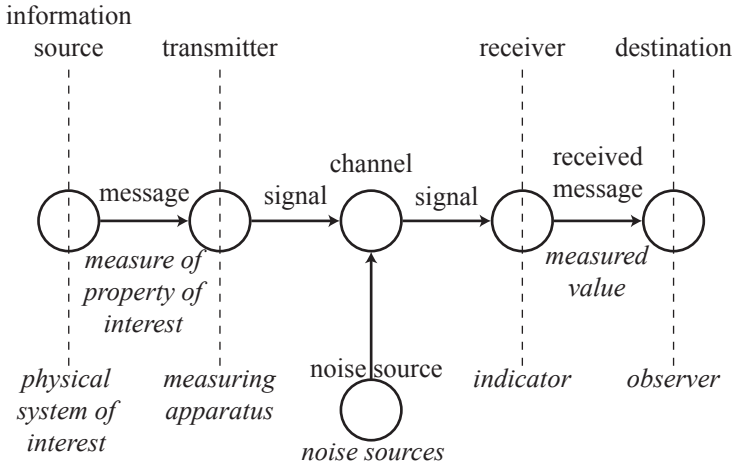


Fig. 22.10 Analogy between communication and a measurement process. Labels in normal text correspond to concepts of communication theory, italic labels to concepts of the measurement process in physics (Rothstein, 1951).

the transmission channel. In our case, this might be a mobile phone that can be used to convert speech containing the information into an electromagnetic wave that can be transmitted through space (the channel). In both cases, noise may degrade the signal as it is transmitted through the channel and reduce the mutual entropy. Eventually the signal will be received, decoded, and displayed. In the communication system this will eventually be the mobile phone of the receiving person which decodes the incoming electromagnetic signals and transforms them into sound waves again. In modern measurement systems the receiver may eventually be an analog-to-digital converter card in a computer system, which allows us to display the measured value. Eventually the message will be received by the destination, i.e., the person listening to the sounds coming out of his mobile phone, who tries to understand what is being said. Equivalently, in physical laboratories there will be an observer, i.e., a physicist reading the measured value and trying to make sense of it. This analogy between communication and measurements in physics shows that it can be very useful for a physicists to care about the results of information theory and the practical implementations of information processing.

22.2.2 Energy dissipation during bit erasure: Landauer's principle

Using the second law of thermodynamics, Landauer has shown that the erasure of a logical bit in a computer dissipates at least the energy $kT \ln 2$. In order to show this, we consider the physical state of a bit in an (isolated) computer. The bit may have either the logical value 0 or 1, represented by macroscopically distinct states of a thermodynamic system. If we assume that we do not know the logical state of this system, it carries the information entropy 1 bit. In thermodynamic language, the bit system may be in either of two macroscopic states, i.e., it fills

a phase space volume of, say, $2n$ possible microscopic states and has a thermodynamic entropy of $S_{\rm bit}^{(0)} = -k_{\rm B}\ln(2n)$. Independent of the state of this bit, the remaining computer system may be in one of a large number N of microscopic states. For example, the different vibrational states of the atoms in solid material, and the positions and momenta of gas atoms within the computer contribute to N. The thermodynamic entropy of the computer (except the bit) is therefore $S_{\rm comp}^{(0)} = -k_{\rm B}\ln N$. The number of physical microscopic states of the total system (the bit plus the remaining system) is therefore $2nN$, and the total thermodynamic entropy is the sum of the entropies of the two subsystems. This state space is depicted schematically in Fig. 22.11(top).

Erasing the bit means finding a physical process that allows us to set the bit to logical zero, no matter what its logical (or physical) state was initially. If we choose a procedure involving only the reversible laws of classical physics, the time evolution of the whole system is given by the Liouville equation which conserves the phase space volume in time. In our example the conservation of phase space volume implies that no matter how the bit erasure is achieved, after the erasure the phase space volume of the whole system has to be $2nN$. Because the number of logical states accessible for the erased bit has reduced to one, the number of possible physical microscopic states of the bit is now n, and its thermodynamic entropy has reduced to $S_{\rm bit}^{(1)} = -k_{\rm B}\ln n$, i.e., the entropy has decreased by $\Delta S_{\rm bit} = -k_{\rm B}\ln 2$. The remaining computer system must therefore fill the phase space of volume $2N$ as shown in Fig. 22.11(bottom). This corresponds to an increase of the entropy of the remaining computer system by $\Delta S_{\rm comp} = k_{\rm B}\ln 2$. We can see from these considerations that the entropy of the whole system remains constant.

Now assume that the computer system (excluding the bit) acts as a heat bath of temperature T. The increase of its entropy means that its energy has increased by

$$\Delta E = T\Delta S = k_{\rm B}T\ln 2.$$

This statement is called Landauer's principle (Landauer, 1961; Landauer, 1993): Erasing a logical bit of information represented by a physical system means the dissipation of the energy $k_{\rm B}T\ln 2$ corresponding to the entropy increase $k_{\rm B}\ln 2$ caused by the erasure.

This value of energy expenditure is commonly referred to as the principal lower limit of energy dissipation in computing. We can see from the above line of argument that it is no problem to dissipate more energy during bit erasure, e.g., by choosing an irreversible physical process which requires an extension of Liouville's equation.

Modern computer machines operate at power dissipation levels orders of magnitude above Landauer's limit.

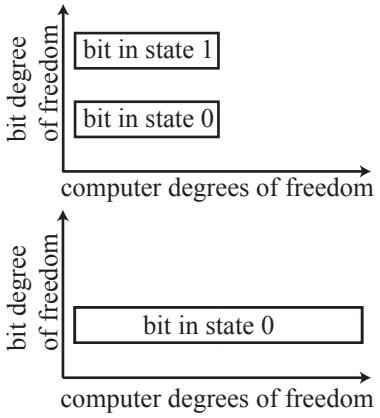


Fig. 22.11 Phase space of the states of a computer with one physical bit in some state (top), and with the same bit erased, i.e., in a well-defined state 0. Conservation of phase space volume by the Liouville-equation the phase space occupied by the computer increases by a factor of two when the bit is erased.

22.2.3 Boolean logic

The information theory described above is based on bits with the two states zero and one. Information is encoded in chains of bits (registers).

The information is processed using logical operations taking bits as the input and producing bits as the output. Present-day computers use boolean logic.[2] In boolean logic we distinguish operations acting on single input bits and those acting on two input bits.

Single-bit operations. There are two operations acting on single bits. This are the functions IDENTITY and NOT. The action of logical operations can be represented in the form of truth tables without referring to the physical implementation of the logic element. For the NOT operation with the input bit i and the output bit j the truth table is shown in Table 22.3. Mathematically we can write IDENTITY$x = x$ and NOT$x = 1 - x$, where $x = 0, 1$. The truth table of a logical operation tells us which elements of the transition matrix $p(j|i)$ are zero and which are one.

Table 22.3 Truth table of the logical NOT operation.

i	j
0	1
1	0

Alternatively, we can use the vector notation for classical bits introduced in section 22.1.2. The identity operation can then be represented by the identity matrix

$$\text{IDENTITY} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{22.9}$$

and the NOT operation is represented by Pauli's σ_x matrix

$$\text{NOT} \equiv \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{22.10}$$

Two-bit operations. Two-bit operations (like the logical AND that we have discussed earlier when introducing the notion of information-related quantities) map the values of two input bits i and j onto one output bit k. There are four possible combinations of the input bits: 00, 01, 10 and 11. The truth table of the AND operation (mathematically: iAND$j = ij$) is shown in Table 22.4.

Table 22.4 Truth table of the logical AND operation.

ij	k
00	0
01	0
10	0
11	1

The first column of this table is the same for all two-bit operations. The second column consists of four bit values. As a consequence, $2^4 = 16$ different two-bit operations with two input and one output bit are conceivable. Regarding the second column of the truth table as a binary number, we can label each possible two-bit operation with this number ranging from 0 to 15. The AND operation has the number $0001 = 1$. The logical OR (mathematically: xOR$y = x + y - xy$) with the truth table shown in Table 22.5 has the number $0111 = 7$. The operation NAND (mathematically: xNAND$y = 1 - xy$) produces $1110 = 14$, the function XOR $0110 = 6$, NOR is $1000 = 8$, and EQUALS is $1001 = 9$.

Table 22.5 Truth table of the logical OR operation.

ij	k
00	0
01	1
10	1
11	1

All possible logical one- and two-bit operations can be realized using only the operations NOT, AND, and OR. Even NOT and AND is sufficient, because

$$x\text{OR}y = \text{NOT}[(\text{NOT}x)\text{AND}(\text{NOT}y)].$$

[2] after George Boole, 1815–1864, English mathematician

We also find, for example,

$$\begin{aligned} \mathrm{NOT}x &= x\mathrm{NAND}x = x\mathrm{NAND}\,1 \\ x\mathrm{AND}y &= (x\mathrm{NAND}y)\mathrm{NAND}(x\mathrm{NAND}y) \\ x\mathrm{OR}y &= (x\mathrm{NAND}x)\mathrm{NAND}(y\mathrm{NAND}y). \end{aligned}$$

Indeed, it is found that all possible one- and two-bit operations can be produced by implementing only the NAND operation.

There are (logically) reversible and (logically) irreversible operations. Logical reversibility is not automatically identical to thermodynamic reversibility. The NOT operation, for example, is logically reversible, because one can infer from any output z the input x. This implies that the loss of information L in a deterministic NOT gate is zero. As we have seen, this is not the case for the logical AND operation which is logically irreversible and a finite information loss L occurs.

A certain complete set of logical operations allows the construction of a universal computer ('Turing machine'). A universal computer can, in principle, compute everything that is computable. Present-day computers are in principle special implementations of this universal computer in which the logical operations are implemented using transistors within the so-called CMOS technology (CMOS-technology means Complementary Metal Oxide Semiconductor technology).

22.2.4 Reversible logic operations

It was mentioned in section 22.1.2 that reversible information processing is the natural classical counterpart of quantum information processing. We therefore introduce a few key concepts of reversible logical operations here.

It is possible to find a complete set of *reversible* logical operations, i.e., operations for which the information loss $L = 0$, such that the input state can be inferred uniquely from the output state. Reversible operations have the same number of input states and output states. If we have n input bits there are 2^n input states linked to the same number of output states. There are $2^n!$ n-bit operations.

One-bit operations. According to the above reasoning there must be two one-bit operations. These are the operations NOT and IDENTITY introduced before. We mentioned that they can be represented by matrices acting on state vectors of the input bit.

Two-bit operations. There are 24 reversible two-bit operations. An important example is the CONTROLLED NOT which has the truth table shown in Table 22.6. The second bit of the output is identical to the output of XOR operating on the input, the first bit of the output is identical to the first bit of the input. The first input bit can be regarded as the control bit. If it is one, then the NOT operation is applied to the

Table 22.6 Truth table of the controlled not operation.

in	out
00	00
01	01
10	11
11	10

second input bit delivering the second output bit, but if it is zero, the second input bit is simply copied to the second output bit.

The action of the CONTROLLED NOT on a two-bit state vector can be described by a 4×4 matrix. If we denote the four two-bit states as $|00\rangle$, $|01\rangle$, $|10\rangle$, and $|11\rangle$, the matrix representation of the controlled not is

$$\text{CONTROLLED NOT} \equiv \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{22.11}$$

Another operation acting on two bits which is important in reversible computing is the SWAP operator which changes $|01\rangle$ to $|10\rangle$, but leaves $|00\rangle$ and $|11\rangle$ unchanged. It has the matrix representation

$$\text{SWAP} \equiv \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

It has been found that the operations NOT, CONTROLLED NOT and CONTROLLED CONTROLLED NOT are a complete set of operations which allow us to construct all possible reversible logical operations. The truth table of the CONTROLLED CONTROLLED NOT is shown in Table 22.7. The first two input bits are the control bits that are copied into the first two output bits. The third bit at the input is only inverted, if both control bits are one.

Table 22.7 Truth table of the CONTROLLED CONTROLLED NOT operation.

in	out
000	000
001	001
010	010
011	011
100	100
101	101
110	111
111	110

It has been theoretically proven that reversible operations can be physically implemented in such a way that no energy is dissipated (Bennett, 1979) implying that computing without energy dissipation is possible. The time evolution of such a computer would then correspond to the time evolution of a conservative physical system. In principle, a universal computer can be built from reversible logical operations (Bennett, 1979; Toffoli, 1981; Fredkin and Toffoli, 1982; Bennett, 1982).

22.3 Brief survey of the theory of quantum information processing

22.3.1 Quantum information theory: the basic idea

The fact that transistors switch smaller and smaller numbers of electrons while operating at faster and faster rates, that we communicate with larger and larger bandwidths over optical fibres, that we store larger and larger numbers of bits in a given volume, poses questions about the fundamental physical limits of these developments. Some of the most cited statements made in this context stem from R. Feynman, for example, 'There's plenty of room at the bottom', the title of his talk for the 1959 annual meeting of the American Physical Society at Caltech

(Feynman, 1992). He made clear that there is no fundamental limit prohibiting us from realizing individual bits of data storage with individual atoms. At another place he stated that '... it seems that the laws of physics present no barriers to reducing the size of computers until bits are the size of atoms and quantum behavior holds dominant sway.' (Feynman, 1985). If we follow the ideas of communication discussed in the previous chapter, it is obvious to ask whether it would be possible to communicate using individual photons, or electrons. We could envisage, for example, encoding a single classical bit of information into the polarization state of the photon, or into the spin state of the electron. We will return to these examples later on.

In general, quantum information processing makes use of our ability to control the coherent unitary time evolution of quantum states. However, this coherent evolution is very fragile and is subject to decoherence as a result of coupling to environmental degrees of freedom, in marked contrast to the time evolution of the states of classical information processing machines which are usually very robust against environmental disruptions. In order to avoid decoherence as much as possible, quantum systems useful as building blocks for information processing are usually small, avoid irrelevant internal degrees of freedom, and minimize coupling to the environment. Small systems have the virtue of exhibiting large energy scales for excitations, which facilitates isolating them dynamically. Furthermore, error correction schemes for quantum information processing have been developed which allow us to correct for disruptions due to the remaining undesired coupling, if it is small enough. Quantum information processing is the extension of reversible classical computing, because dissipation of energy and decoherence are related.

The motivation for developing information processing with quantum systems is that quantum algorithms can solve some information processing tasks of practical interest much more efficiently than today's classical algorithms. The two most prominent examples are Grover's search algorithm, and Shor's prime number factorization algorithm, the latter being a major threat for the security of the widely used RSA[3] method of data encryption. Essentially, quantum mechanics enhances our possibilities for processing digital information.

In general, any quantum mechanical two-level system may provide us with the possibility of encoding a classical bit. The term *qubit* was therefore coined as the quantum mechanical analogue of classical two-state systems holding one bit of information. In contrast to a classical bit, a qubit is a quantum mechanical system that can consist of a very small number of particles. Therefore, a thermodynamic description is typically not appropriate. Data storage is achieved by selecting two particular quantum states to represent one bit of information. In contrast to a classical bit, the qubit can be in a superposition of these two quantum

[3]Named after R. Rivest, A. Shamir, and L. Adleman, who invented the scheme in 1977.

states, i.e., in a superposition of zero and one.

According to the rules of quantum mechanics, measurements of qubits (reading the qubit) do necessarily mean that we change the qubit significantly, and only probabilistic predictions about the result of the measurement can be made.

On a more abstract level we stated on page 474 two basic ingredients for the description of the properties of classical information. By analogy we may require for quantum information systems:

(1) A Hilbert space spanned by orthonormal quantum states $|\,n\rangle$. The state of a particular system at a given instant is, for example, described by a density matrix $\hat{\rho}$.

(2) A probability distribution p_n which allows us to write the state of the system as $\hat{\rho} = \sum_n p_n \,|\,n\rangle\,\langle n|$.

Seen from the quantum information perspective, the density matrix of a quantum system describes our uncertainty about the state of the system, as the probability distribution does in a classical statistical system.

22.3.2 Qubits

The smallest Hilbert space suitable for information storage is spanned by two orthogonal quantum states. Such a system represents the abstract realization of one qubit of information. Of course, the question arises as to how the information stored in one qubit can be read by measurement. It is immediately clear for us physicists that measurement of a quantum system is very much different from measurement of a classical system, as it usually completely changes the quantum state of the system being measured. Before we go into more detail about how quantum information differs from classical information, we summarize the way qubits are represented in our quantum mechanical language.

The qubit is represented by the superposition of two orthogonal states, i.e., by the general state of a two-level system. Examples for two-level systems in physics are the electron spin ($\uparrow$, $\downarrow$), the polarization state of a photon (right or left circularly polarized), or two electronic energy levels of an atom (ground state, excited state). In the following, we will briefly introduce different notations that are commonly used to describe qubits.

Dirac notation. In general, the state of a qubit can be written in *Dirac notation* as

$$|\,\psi\rangle = \alpha_0\,|\,0\rangle + \alpha_1\,|\,1\rangle\,, \tag{22.12}$$

where normalization requires

$$\alpha_0^2 + \alpha_1^2 = 1. \tag{22.13}$$

The qubit may be interpreted as a coherent superposition of the two states 0 and 1 of a classical bit. Indeed, the two state vectors of a classical bit can be formally seen as a small subset of the two qubit

states (either $\alpha_0 = 1$, $\alpha_1 = 0$, or $\alpha_0 = 0$, $\alpha_1 = 1$). However, the physical implementations of the two are vastly different.

The normalization condition (22.13) reduces the number of independent qubit parameters from four (real and imaginary parts of the two complex numbers α_0 and α_1) to three. In addition, the absolute phase of the wave function is arbitrary such that we can always choose α_0 to be real, and reduce the number of relevant parameters to two. One way to parametrize the basis states, such that the normalization condition is automatically met, is

$$\alpha_0 = \cos\frac{\theta}{2} \qquad \alpha_1 = e^{i\delta}\sin\frac{\theta}{2}. \tag{22.14}$$

Here θ parametrizes the probabilities $p_{0/1}$ of finding the system in one or other state ($p_0 = \cos^2\theta/2$ and $p_1 = \sin^2\theta/2$). The parameter δ is the relative phase of the two states.

Systems with spin 1/2, such as the spin of an individual electron, are natural realizations of two-level quantum systems. However, any other two-level quantum system can be described in exactly the same way. We will use the electron spin as a paradigmatic example for introducing ways of describing qubits alternative to the Dirac notation in eq. (22.12).

A system of two qubits (we label them A and B) can be described with state vectors in a Hilbert space spanned by four basis states. For example, we can choose them to be (in Dirac notation)

$$|00\rangle\,, |01\rangle\,, |10\rangle\,, |11\rangle\,,$$

where $|ij\rangle = |i\rangle_{\mathrm{A}} \otimes |j\rangle_{\mathrm{B}}$. The state of two classical bits would be exactly one of these four orthogonal states. Systems of two qubits can be in a state described by any of the linear combinations

$$|\psi\rangle = \alpha_0|00\rangle + \alpha_1|01\rangle + \alpha_2|10\rangle + \alpha_3|11\rangle\,,$$

where the coefficients are complex numbers obeying the normalization condition

$$|\alpha_0|^2 + |\alpha_1|^2 + |\alpha_2|^2 + |\alpha_3|^2 = 1.$$

As for single qubits the overall phase of the wave function is of no significance, and one of the coefficients can be taken to be real-valued. The remaining seven parameters (four absolute values and three phases) are not independent as a result of the normalization condition. The wave function is therefore determined by six real parameters, i.e., more than the four parameters required for describing the states of two independent (uncorrelated) qubits. The reason is that two qubits can be correlated, or entangled.

Two qubits are called *entangled* if their wave function cannot be written as the product of two single qubit states. Examples of entangled states are

$$\begin{aligned} |\psi\rangle &= \frac{1}{\sqrt{2}}\left(|00\rangle \pm |11\rangle\right) \\ |\psi\rangle &= \frac{1}{\sqrt{2}}\left(|01\rangle \pm |10\rangle\right). \end{aligned}$$

In contrast, the following states do not describe entangled qubits:

$$\begin{aligned} |\psi\rangle &= \frac{1}{\sqrt{2}}\left(|00\rangle \pm |01\rangle\right) = |0\rangle_{\mathrm{A}} \otimes \frac{1}{\sqrt{2}}\left(|0\rangle_{\mathrm{B}} \pm |1\rangle_{\mathrm{B}}\right) \\ |\psi\rangle &= |01\rangle = |0\rangle_{\mathrm{A}} \otimes |1\rangle_{\mathrm{B}} \end{aligned}$$

General n-qubit states (n is integer) can be written in Dirac notation as a linear combination of the 2^n basis states $|i\rangle_n$, where the numbers i are n-digit binary numbers (i.e., they may have trailing zeros) obeying $0 \leq i < 2^n$. For example, one state of the eight three-qubit basis states could be written as $|5\rangle_3$, or equivalently as $|101\rangle$; another one would be $|2\rangle_3 \equiv |010\rangle$. With this notation, a general n-qubit state can be written as

$$|\psi\rangle = \sum_{i=0}^{2^n-1} \alpha_i |i\rangle_n ,$$

where the coefficients α_i obey the normalization condition

$$\sum_{i=1}^{2^n} |\alpha_i|^2 = 1.$$

Pauli notation. In the *Pauli notation*, the two orthogonal states $|0\rangle$ and $|1\rangle$ of a single qubit are written as spinors in vector notation. Using the parametrization for a single qubit state introduced above, we can write an arbitrary qubit state as the two-component spinor

$$|\psi\rangle \equiv \begin{pmatrix} \cos\theta/2 \\ e^{i\delta}\sin\theta/2 \end{pmatrix}.$$

Some special states are given in Table 22.8.

Like the single-qubit states, two-qubit states can be written in spinor notation by writing the coefficients of the four orthonormalized basis

Table 22.8 Special qubit states.

θ	δ	state	polarization vector
0	-	$\|0\rangle$	$(0,0,1)$
π	-	$\|1\rangle$	$(0,0,-1)$
$\frac{\pi}{2}$	0	$\frac{1}{\sqrt{2}}\left(\|0\rangle + \|1\rangle\right)$	$(1,0,0)$
$\frac{\pi}{2}$	π	$\frac{1}{\sqrt{2}}\left(\|0\rangle - \|1\rangle\right)$	$(-1,0,0)$
$\frac{\pi}{2}$	$\frac{\pi}{2}$	$\frac{1}{\sqrt{2}}\left(\|0\rangle + i\|1\rangle\right)$	$(0,1,0)$
$\frac{\pi}{2}$	$-\frac{\pi}{2}$	$\frac{1}{\sqrt{2}}\left(\|0\rangle - i\|1\rangle\right)$	$(0,-1,0)$

functions in vector form as

$$|\psi\rangle = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix},$$

and a general n-qubit state would be

$$|\psi\rangle = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_N \end{pmatrix},$$

where $N = 2^n - 1$.

Polarization vector of a single qubit and Bloch sphere. A two-level quantum state (i.e., a single qubit) may alternatively be uniquely described by its *polarization vector* $\mathbf{P} = (P_x, P_y, P_z)$. Its components are defined as the expectation values of Pauli's spin matrices σ_x, σ_y, σ_z [see eq. (3.14)], i.e., $P_i := \langle \sigma_i \rangle$. Expressing the components P_i using the parametrization (22.14), we obtain

$$\begin{aligned} P_x &= \sin\theta \cos\delta \\ P_y &= \sin\theta \sin\delta \\ P_z &= \cos\theta. \end{aligned}$$

This is the parametrization of a three-dimensional vector of length 1 ($|\mathbf{P}| = 1$). We can represent this vector in a three-dimensional coordinate system by putting its starting point at the origin. Its end point will be somewhere on the surface of a unit sphere as depicted in Fig. 22.12. The polarization vector encloses the angle θ with the z-axis, and the angle δ represents the azimuth. This representation of a qubit state on the surface of a unit sphere is called the *Bloch sphere representation*. The direction of the polarization for special states has been included in Table 22.8 and is shown in Fig. 22.12. When applying this representation in practice it is very important to remember that orthogonal quantum states are represented here as *antiparallel* polarization vectors rather than orthogonal vectors.

There is no commonly used representation of the states of two or more qubits which is equivalent to the polarization vector notation, and also, the Bloch-sphere visualization of single qubit states has no commonly used generalization for many qubits.

Density matrix notation. The density matrix representing the single qubit state (22.12) is given by

$$\begin{aligned} \hat{\rho} := \begin{pmatrix} |\alpha|^2 & \alpha\beta^\star \\ \alpha^\star\beta & |\beta|^2 \end{pmatrix} &= \begin{pmatrix} \cos^2\frac{\theta}{2} & \frac{1}{2}e^{-i\delta}\sin\theta \\ \frac{1}{2}e^{i\delta}\sin\theta & \sin^2\frac{\theta}{2} \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} 1+P_z & P_x - iP_y \\ P_x + iP_y & 1-P_z \end{pmatrix}. \end{aligned} \quad (22.15)$$

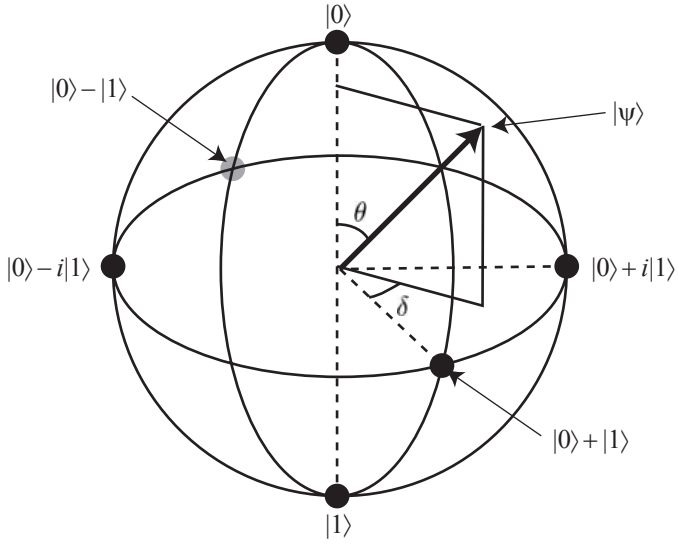


Fig. 22.12 Bloch sphere representation of a qubit.

This density matrix uniquely represents a pure qubit state. It can, for example, be used to calculate the expectation value of any arbitrary observable $\hat{O}$ via

$$\left\langle \hat{O} \right\rangle = \mathrm{tr}\left[\hat{O}\hat{\rho}\right].$$

The diagonal elements of the density matrix are the probabilities $p_{0/1}$ to find the qubit in state $|0\rangle$ or $|1\rangle$. The off-diagonal elements are called *interferences*.

The virtue of the density matrix representation is that it can not only be used to describe pure quantum states, but it is also well suited to describing statistical mixtures. The definition of uncertainty about a quantum state and therefore the notion of quantum information is also based on the density matrix notation. In quantum information theory, the density matrix is interpreted as a mathematical description of our uncertainty about a quantum system.

An arbitrary composite system made of the two subsystems A and B can be described on the basis of the product basis

$$|i\alpha\rangle = |i\rangle \otimes |\alpha\rangle ,$$

and the general wave function of the composite system can be written in Dirac notation as

$$|\psi\rangle = \sum_{i\alpha} a_{i\alpha} |i\alpha\rangle .$$

Correspondingly, the expectation value of an arbitrary operator $\hat{O}$ is

given by

$$\begin{aligned}\left\langle \hat{O} \right\rangle &= \left(\sum_{i\alpha} a^{\star}_{i\alpha} \left\langle i\alpha \right| \right) \hat{O} \left(\sum_{j\beta} a_{j\beta} \left| j\beta \right\rangle \right) \\ &= \sum_{i\alpha} \sum_{j\beta} \left\langle i\alpha \right| \hat{O} \left| j\beta \right\rangle a_{j\beta} a^{\star}_{i\alpha} \\ &= \text{trace} \left(\hat{O}\hat{\rho} \right),\end{aligned}$$

where the elements of the density matrix $\hat{\rho}$ have been defined using the coefficients of the basis functions, i.e.,

$$\rho_{j\beta,i\alpha} = a_{j\beta} a^{\star}_{i\alpha}.$$

Again, the diagonal elements of the density matrix are the probabilities of finding the system in the corresponding basis states. The off-diagonal matrix elements are the interferences.

We consider a few density matrices of two-qubit states as examples. The density matrix of the state $\left| 01 \right\rangle$ is

$$\hat{\rho} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

For the state

$$\left| \psi \right\rangle = \frac{1}{\sqrt{2}} \left(\left| 00 \right\rangle \pm \left| 01 \right\rangle \right) = \left| 0 \right\rangle_{\text{A}} \otimes \frac{1}{\sqrt{2}} \left(\left| 0 \right\rangle_{\text{B}} \pm \left| 1 \right\rangle_{\text{B}} \right),$$

the density matrix is

$$\hat{\rho} = \begin{pmatrix} 1/2 & \pm 1/2 & 0 & 0 \\ \pm 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The entangled state

$$\left| \psi \right\rangle = \frac{1}{\sqrt{2}} \left(\left| \uparrow\uparrow \right\rangle \pm \left| \downarrow\downarrow \right\rangle \right)$$

has the density matrix

$$\hat{\rho} = \begin{pmatrix} 1/2 & 0 & 0 & \pm 1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \pm 1/2 & 0 & 0 & 1/2 \end{pmatrix}.$$

Reduced density matrix. A special case arises if the operator $\hat{O}$ acts only on subsystem A. For example, if the composite system consists of two electrons that are used as spin-qubits, a measurement on only one of the two spins would be represented by such an operator. The expectation value is then given by

$$\begin{aligned}\left\langle \hat{O} \right\rangle &= \left(\sum_{i\alpha} a^{\star}_{i\alpha} \langle i\alpha| \right) \hat{O} \left(\sum_{j\beta} a_{j\beta} |j\beta\rangle \right) \\ &= \sum_{i\alpha} \sum_{j\beta} \langle i| \hat{O} |j\rangle \, \delta_{\alpha\beta} a_{j\beta} a^{\star}_{i\alpha} \\ &= \sum_{ij} \langle i| \hat{O} |j\rangle \sum_{\alpha\beta} \delta_{\alpha\beta} \rho_{j\beta,i\alpha} \\ &= \sum_{ij} \langle i| \hat{O} |j\rangle \sum_{\alpha} \rho_{j\alpha,i\alpha}.\end{aligned}$$

We now define the elements of the reduced density matrix, i.e., the density matrix of the subsystem with the basis vectors $|i\rangle_{\mathrm{A}}$ as

$$\rho_{ij} = \sum_{\alpha} \rho_{j\alpha,i\alpha},$$

i.e., we calculate the partial trace of the density matrix of the composite system (we trace out subsystem B). Using this definition of the reduced density matrix, we again find

$$\left\langle \hat{O} \right\rangle = \mathrm{trace}\left(\hat{O}\hat{\rho} \right).$$

The reduced density matrix allows us to calculate all properties that can be retrieved from measurements on subsystem A alone. It therefore constitutes a complete description of subsystem A, for which an equivalent wave function representation does not necessarily exist.

We consider again a few examples of two-qubit states. For the state $|01\rangle$ the reduced density matrix is given by

$$\hat{\rho} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

For the state

$$|\psi\rangle = \frac{1}{\sqrt{2}} \left(|00\rangle \pm |01\rangle \right) = |0\rangle_{\mathrm{A}} \otimes \frac{1}{\sqrt{2}} \left(|0\rangle_{\mathrm{B}} \pm |1\rangle_{\mathrm{B}} \right),$$

the reduced density matrix is again

$$\hat{\rho} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

telling us that a measurement on subsystem A is not sufficient to distinguish the two two-qubit states.

The entangled states

$$|\psi\rangle = \frac{1}{\sqrt{2}}\left(|00\rangle \pm |11\rangle\right)$$

have the reduced density matrix

$$\hat{\rho} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

The sign $\pm$ obviously does not play a significant role for observations in subsystem A, and the two different two-qubit states cannot be distinguished by observations of subsystem A alone.

General properties of density matrices. In general, density matrices have the following three properties:

(1) they are self-adjoint, i.e., $\rho_{ij} = \rho^{\star}_{ji}$,
(2) their trace is 1, i.e., trace$\hat{\rho} = 1$,
(3) all their eigenvalues are larger than or equal to zero.

If we have a composite system consisting of subsystems A and B, entanglement in the state of the composite system can be detected by calculating the reduced density matrix for subsystem A. If the two subsystems A and B are not entangled, the reduced density matrix obeys

$$\hat{\rho}^2 = \hat{\rho},$$

and it represents a pure state that can also be represented by a wave function. If the two subsystem A and B are entangled,

$$\hat{\rho}^2 \neq \hat{\rho},$$

and it represents a mixed state that cannot be represented by a single wave function, but by a statistical mixture of wave functions.

22.3.3 Qubit operations

In quantum mechanics, unitary transformations govern the time evolution of physical systems. If we wish to process information encoded in individual qubits, we have to control the unitary time evolution of systems of qubits. As in classical information processing we distinguish operations acting on only a single qubit (one-qubit gates), on pairs of qubits (two-qubit gates), and on more than two qubits. Quantum information theory teaches us that arbitrary unitary transformations on n qubits can be approximated (with arbitrary precision) by applying a sequence of one- and two-qubit gates (DiVincenzo, 1995; Barenco *et al.*, 1995). This is in analogy to classical information processing where an arbitrary computation can be broken down into a sequence of operations acting only on single classical bits or two classical bits. As a consequence, attempts to implement quantum computation can concentrate on the realization of single-qubit and two-qubit gates.

Given a particular single-qubit state represented as a vector ending on the Bloch sphere, a general unitary transformation would allow us to rotate this vector to an arbitrary other position on the Bloch sphere. Operations on two qubits usually invoke interactions between them and lead to control over entanglement. Arbitrary single-qubit rotations together with the implementation of a CONTROLLED NOT gate (22.11) for two qubits are sufficient to perform any quantum computation.

A general single-qubit rotation is described by the 2×2-matrix

$$U_1(\alpha, \beta, \gamma) = \begin{pmatrix} e^{i(\beta+\gamma)/2}\cos(\alpha/2) & e^{-i(\beta-\gamma)/2} i \sin(\alpha/2) \\ e^{i(\beta-\gamma)/2} i \sin(\alpha/2) & e^{-i(\beta+\gamma)/2}\cos(\alpha/2) \end{pmatrix},$$

where α, β, and γ are real parameters (Euler angles).

The classical reversible operations on a single qubit have their quantum analogues. The IDENTITY operation represented by the 2×2 unity matrix (22.9) is an operation that can also be applied to a qubit leaving its state unchanged. Considering the above general single-qubit rotation, it corresponds to $U_1(0, 0, 0)$. The matrix (22.10) representing the classical NOT which acts on an arbitrary qubit like

$$\sigma_x \,|\,\psi\rangle = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \beta \\ \alpha \end{pmatrix}$$

is called a qubit flip and corresponds to $-iU_1(\pi, \pi/2, \pi/2)$.

Of course there are a large number of other single-qubit operations without a classical analogue. For example, the Pauli matrix σ_z acts as

$$\sigma_z \,|\,\psi\rangle = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha \\ e^{i\pi}\beta \end{pmatrix},$$

which is a π-phase shift operation equivalent to $-iU_1(0, \pi/2, \pi/2)$. The Pauli matrix σ_y is a combined qubit flip and π-phase shift operation:

$$\sigma_y \,|\,\psi\rangle = i\sigma_x\sigma_z \,|\,\psi\rangle = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} -i\beta \\ i\alpha \end{pmatrix},$$

corresponding to $-iU_1(\pi, \pi/2, -\pi/2)$. Another single-qubit operation that is frequently encountered in quantum information processing is the *Walsh–Hadamard transformation*,

$$H \,|\,\psi\rangle = \frac{1}{\sqrt{2}}(\sigma_x + \sigma_z) \,|\,\psi\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \alpha + \beta \\ \alpha - \beta \end{pmatrix},$$

corresponding to $-iU_1(\pi/2, \pi/2, \pi/2)$. Applied to the qubit state $|\,0\rangle$ it gives the superposition $(|\,0\rangle + |\,1\rangle)/\sqrt{2}$.

22.4 Implementing qubits and qubit operations

The challenge for experimental physicists trying to implement quantum information processing schemes is to design tailored two-level quantum

systems that can be prepared in a well-defined initial state, then coherently manipulated with the desired one- or two-qubit gates, and then measured in order to learn the result of the computation. The coherent manipulation stage requires that the time evolution is governed by a well-known and well-controllable time-dependent hamiltonian $\hat{H}(t)$ which brings about the desired unitary transformation of the prepared initial state, usually by controlling time-dependent external parameters. In particular, this latter aspect has inspired experimentalists in recent years to perform novel experiments, a few of which we will discuss in the following.

22.4.1 Free oscillations of a double quantum dot charge qubit

We start the discussion of possible realizations of single-qubit gates by showing an experiment performed on a double quantum dot system which can be interpreted as a so-called *charge qubit*. The experiment was performed by T. Hayashi and coworkers in 2003 (Hayashi *et al.*, 2003). Figure 22.13(a) shows the double quantum dot structure which has been laterally defined based on a heterostructure with a two-dimensional electron gas. A channel of 500 nm width was defined by wet-chemical etching. The three gates G_{L}, G_{C}, and G_{R} were fabricated right above the channel in order to allow the formation of tunneling barriers controlled

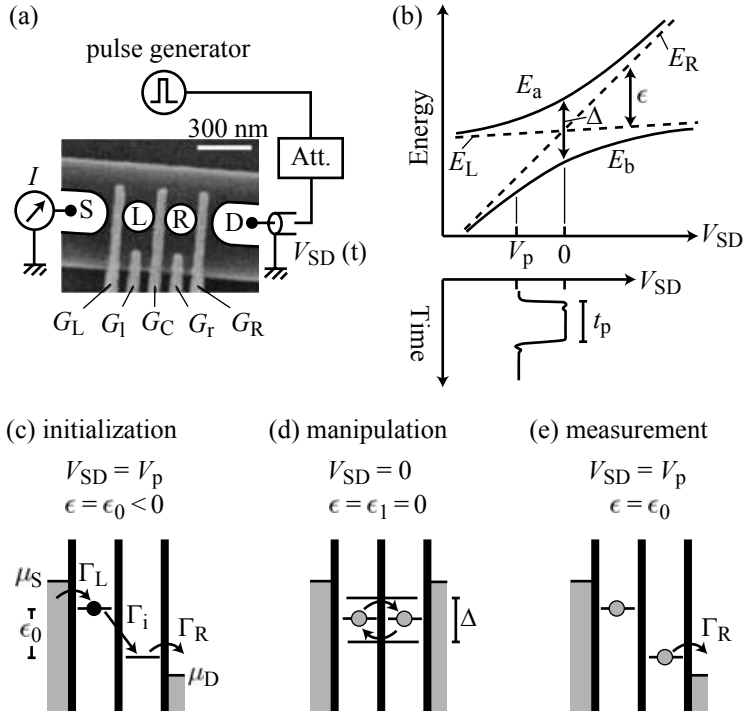


Fig. 22.13 (a) Schematic representation of the measurement setup combined with an SEM image of the lateral double dot structure. (b) Energy level diagram with 'bonding' and 'antibonding' branch resulting from the avoided crossing of two levels. (c)–(e) Energy-level diagram during qubit initialization (c), the time of coherent oscillation (d) and the read-out measurement (e). (Reprinted with permission from Hayashi *et al.*, 2003. Copyright 2003 by the American Physical Society.)

by the voltages on these gates. Two additional gates G_l and G_r serve as plunger gates for the quantum dots L and R. For the experiment, the central gate G_C controlling the coupling between the dots, and the plunger gates G_l and G_r were adjusted such that at zero source–drain voltage a level in dot L and another level in dot R are degenerate and split into a symmetric and an antisymmetric state as a result of the tunneling coupling [see Fig. 22.13(d)]. Applying a finite source–drain voltage lifts this degeneracy, the two levels separate in energy, and the corresponding wave functions separate in space [Fig. 22.13(b), (d)].

The two basis states of the qubit are taken to be the two charge states $|\,L\rangle \equiv |\,0\rangle$ and $|\,R\rangle \equiv |\,1\rangle$ with the additional electron in the left dot or in the right dot, respectively. The time evolution of the qubit system is governed by the hamiltonian matrix

$$\hat{H}(t) = \begin{pmatrix} \epsilon_\mathrm{L}(t) & t \\ t^\star & \epsilon_\mathrm{R}(t) \end{pmatrix}.$$

The energy levels of the additional electron in the left and right dot depend on time, because they can be shifted by applying suitable time-dependent gate or source–drain voltages.

The system is initialized by applying a suitable source–drain voltage for a suitable time such that an electron can tunnel from the source contact into the state $|\,L\rangle$ [Fig. 22.13(c)]. Then the source–drain voltage is abruptly (nonadiabatically) switched to zero for a well-defined time span t_p [Fig. 22.13(d)]. During switching the electron remains in the state $|\,L\rangle$ which is, however, no longer an eigenstate of the system. In the new configuration the eigenstates of the double dot system are the symmetric state $|\,S\rangle$ and the antisymmetric state $|\,A\rangle$ given by

$$\begin{aligned} |\,S\rangle &= \frac{1}{\sqrt{2}}\,(|\,L\rangle + |\,R\rangle\,) \\ |\,A\rangle &= \frac{1}{\sqrt{2}}\,(|\,L\rangle - |\,R\rangle\,). \end{aligned}$$

The state of the system immediately after switching is therefore given by

$$|\,\psi(t=0)\rangle = |\,L\rangle = \frac{1}{\sqrt{2}}\,(|\,S\rangle + |\,A\rangle\,),$$

which is a superposition of the two new eigenstates whose energies differ by the symmetric–antisymmetric splitting Δ [Fig. 22.13(b)]. The time evolution of this superposition of states is given by

$$|\,\psi(t)\rangle = \frac{1}{\sqrt{2}}\left(|\,S\rangle + |\,A\rangle\, e^{i\Delta t/\hbar}\right) = \cos\frac{\Delta t}{2\hbar}\,|\,L\rangle + i\sin\frac{\Delta t}{2\hbar}\,|\,R\rangle\,. \quad (22.16)$$

This time evolution is called the free oscillation of the qubit which is periodic with frequency $\Delta/2h$. The motion of the state vector on the Bloch sphere (inset of Fig. 22.14) goes along the meridian given by the time-dependent angles $\theta(t) = \Delta t/\hbar$ and $\delta(t) = \pi/2$. The final state of the system is given by $|\,\psi(t_p)\rangle$.

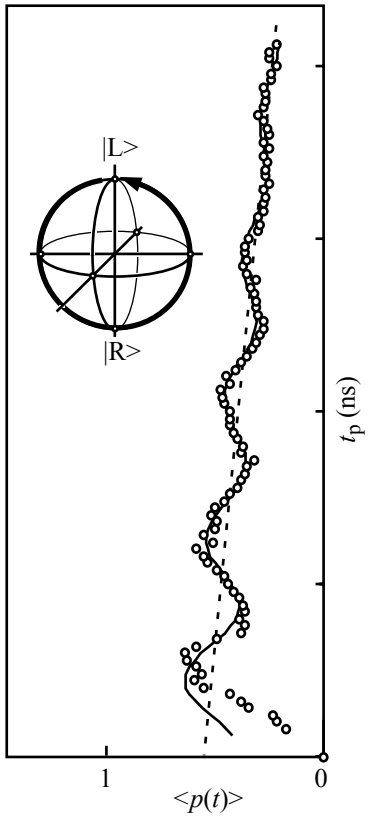


Fig. 22.14 Coherent oscillations of the charge qubit detected by the dc current through the double quantum dot system. (Reprinted from Fujisawa *et al.*, 2004 with permission from Elsevier.)

The read-out of the qubit state after time t_p [Fig. 22.13(e)] is started by switching the system back to finite source–drain voltage. This gives the electron the possibility of tunneling into the drain contact and thereby contribute to the tunneling current. Of course, the current originating from a single electron cannot be measured in this experiment. For this reason, the sequence described above is repeated with a repetition rate $1/t_r$ (in this experiment 100 MHz). If this rate is larger than the rate with which an electron can leak from the left dot into the drain contact, a time-averaged current results which depends on the pulse duration t_p. This current is proportional to the probability that the electron is measured in the right dot

$$\langle I \rangle \propto p(t_p) = \sin^2 \frac{\Delta t_p}{2\hbar}.$$

The qubit state is measured by projection onto the basis state $|R\rangle$. If we could detect individual electrons that have left the right dot with a sensitive charge detector, each cycle would give either a measurement value 'electron detected' or 'no electron detected', with probabilities $p(t)$ and $1-p(t)$, respectively. By averaging over many cycles, the probability $p(t)$ can be measured via the average current $\langle I \rangle$.

If $t_p = (n-1/2)h/\Delta$ (n integer larger than zero), then the final state is $|R\rangle$ and the current should be maximum. These t_p values realize a qubit flip (NOT). If $t_p = nh/\Delta$ the final state is $|L\rangle$ and the current is minimum. These t_p values realize the IDENTITY operation. The current is expected to oscillate as a function of pulse duration t_p with the frequency of the free oscillations.

Figure 22.14 shows these oscillations as measured in the experiment. In contrast to the simple theory of the free qubit oscillations outlined above, the oscillations decay within about 2 ns. The reason is the decoherence of the superposition state. The decoherence time extracted from this experiment is about 0.8 ns.

22.4.2 Rabi oscillations of an excitonic qubit

The states of single qubits can also be rotated by coupling a charge qubit to an external oscillating electric field. In the experimental example to be discussed now (Zrenner *et al.*, 2002), the qubit is formed by two states of a single self-assembled quantum dot. Self-assembled quantum dots made of $In_{0.5}Ga_{0.5}As$ were embedded in a Schottky diode based on GaAs as shown in Fig. 22.15(c). Using a large area top gate with openings of 100 to 500 nm diameter individual quantum dots were selected for the illumination with laser light [see schematic Fig. 22.15(c)]. The laser light can excite Coulomb-coupled electron–hole pairs, also called excitons, if the excitation energy is above band gap, as shown in Fig. 22.15(b). The two states of the corresponding qubit are shown in Fig. 22.15(a). The two states are $|0\rangle$ (no electron–hole pair in the dot), and $|X\rangle$ (one exciton in the dot).

The time evolution of the illuminated system is governed by the hamil-

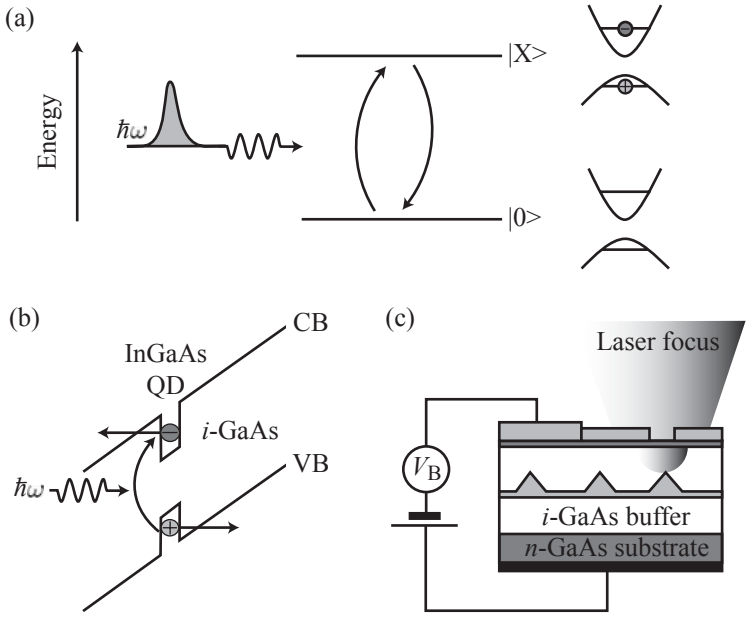


Fig. 22.15 (a) Schematic representation of the experimental system: a coherent laser pulse excites Rabi oscillations between the states $|0\rangle$ and $|X\rangle$. (b) Conduction band edge (CB) and valence band edge (VB) of the structure in growth direction. Photon absorption creates an electron–hole pair (exciton) in the quantum dot. These can leave the dot as a result of the applied electric field. (c) Cross section through the device and electronic setup. (Zrenner *et al.*, 2002. Reprinted by permission from Macmillan Publishers Ltd, copyright 2002.)

tonian

$$\hat{H} = \hat{H}_0 + V(t),$$

where

$$V(t) = -eEx \sin \omega t = -eEx \frac{1}{2i} \left(e^{i\omega t} - e^{-i\omega t} \right).$$

For getting insight into the time evolution of the system, we assume that the time-dependent Schrödinger equation governed by $\hat{H}_0$,

$$i\hbar \partial_t \, |\psi_n(t)\rangle = \hat{H}_0 \, |\psi_n(t)\rangle \, ,$$

is solved by the basis functions

$$|\psi_n(t)\rangle = |n\rangle \, e^{-i\omega_n t},$$

where $|n\rangle$ can be $|0\rangle$ or $|X\rangle$, but also any other possible excitation of the system. We can now expand the solution of the full problem

$$i\hbar \partial_t \, |\psi(t)\rangle = \hat{H} \, |\psi(t)\rangle \tag{22.17}$$

in terms of the time-dependent basis states of the unperturbed problem, i.e.,

$$|\psi(t)\rangle = \sum_n a_n(t) \, |\psi_n(t)\rangle = \sum_n a_n(t) \, |n\rangle \, e^{-i\omega_n t}.$$

Inserting this *Ansatz* into eq. (22.17) we obtain a system of equations for the coefficients $a_n(t)$:

$$i\hbar \partial_t a_m(t) = -\frac{eE}{2i} \sum_n a_n(t) x_{mn} \left(e^{i(\omega_{mn}+\omega)t} - e^{i(\omega_{mn}-\omega)t} \right).$$

Here, $\omega_{mn} = \omega_m - \omega_n$. We now assume that our laser excitation has a frequency ω which is very close to the energy difference $\omega_{\rm fi}$ between the final state $|\,X\rangle$ and the initial state $|\,0\rangle$, i.e.,

$$\omega = \omega_{\rm fi} + \epsilon.$$

We further assume that at time $t < 0$ the system is in the initial state $|\,0\rangle$. The laser excitation is switched on abruptly at time $t = 0$. We now use the so-called *secular approximation*, which means that we only take the coefficients $a_{\rm i}(t)$ and $a_{\rm f}(t)$ into account. This approximation is reasonable as long as the perturbation is small compared to other excitations in the system. This approximation leads to the two equations

$$\begin{aligned} i\hbar\partial_t a_{\rm i}(t) &= -\frac{eE}{2i} a_{\rm f}(t) x_{\rm if} e^{+i\epsilon t} \\ i\hbar\partial_t a_{\rm f}(t) &= +\frac{eE}{2i} a_{\rm i}(t) x_{\rm if}^\star e^{-i\epsilon t}. \end{aligned}$$

This system of equations can be solved exactly. To this end we define

$$a_{\rm i}(t) = e^{i\epsilon t/2} b_{\rm i}(t) \;\text{ and }\; a_{\rm f}(t) = e^{-i\epsilon t/2} b_{\rm f}(t)$$

and obtain

$$\begin{aligned} i\hbar\partial_t b_{\rm i}(t) &= \frac{\hbar\epsilon}{2} b_{\rm i}(t) - \frac{eE}{2i} x_{\rm if} b_{\rm f}(t) \\ i\hbar\partial_t b_{\rm f}(t) &= \frac{eE}{2i} x_{\rm if}^\star b_{\rm i}(t) - \frac{\hbar\epsilon}{2} b_{\rm f}(t). \end{aligned}$$

Solving for the two coefficients gives the harmonic oscillator equation

$$\partial_t^2 b_{\rm i/f}(t) + \Omega^2 b_{\rm i/f}(t) = 0,$$

where

$$\Omega = \sqrt{\left(\frac{eE}{2\hbar}\right)^2 |x_{\rm if}|^2 + \left(\frac{\epsilon}{2}\right)^2}.$$

The initial condition of our problem at time zero is $b_{\rm i}(0) = 1$ and $b_{\rm f}(0) = 0$. Using the system of differential equations, we obtain for the first derivatives $\partial_t b_{\rm i}(0) = \epsilon/2i$ and $\partial_t b_{\rm f}(0) = -eEx_{\rm if}^\star/2\hbar$ and the solution of the problem is

$$\begin{aligned} b_{\rm i}(t) &= \cos\Omega t - \frac{i\epsilon}{2\Omega}\sin\Omega t \\ b_{\rm f}(t) &= \frac{-eEx_{\rm if}^\star}{2\hbar\Omega}\sin\Omega t. \end{aligned}$$

The system is therefore described by the wave function

$$\begin{aligned} |\,\psi(t)\rangle &= a_{\rm i}(t)\,|\,0\rangle\, e^{-i\omega_0 t} + a_{\rm f}(t)\,|\,X\rangle\, e^{-i\omega_{\rm X} t} \\ &= e^{i\epsilon t/2} b_{\rm i}(t)\,|\,0\rangle\, e^{-i\omega_0 t} + e^{-i\epsilon t/2} b_{\rm f}(t)\,|\,X\rangle\, e^{-i\omega_{\rm X} t} \\ &= e^{i\epsilon t/2}\left[\cos\Omega t - \frac{i\epsilon}{2\Omega}\sin\Omega t\right]|\,0\rangle\, e^{-i\omega_0 t} \\ &\quad -e^{-i\epsilon t/2}\left[\frac{eEx_{if}^\star}{2\hbar\Omega}\sin\Omega t\right]|\,X\rangle\, e^{-i\omega_{\rm X} t}. \end{aligned}$$

This describes the so-called *Rabi oscillations* of the excitonic system.

In the experiment, a laser pulse with a finite length t_p is applied to the system. When the laser pulse is abruptly switched off, the system will be in the state $|\psi(t_p)\rangle$. As a result of the internal static electric field in the system [see Fig. 22.15(b)] the exciton can decay by tunneling of the electron and the hole into the contacts, thereby contributing to a photocurrent. This process acts as the read-out measurement of the qubit state. As in the case of the double quantum dot charge qubit, the current caused by a single electron–hole pair cannot be measured. Therefore the laser pulse is repeated with a repetition rate $1/t_\mathrm{r}$. If this rate is larger than the rate for excitonic recombination, but smaller than the tunneling rates, a time-averaged photocurrent results which depends on the pulse duration t_p, but also on the intensity of the laser pulse given by the strength E of the electric field. The time-averaged photocurrent is proportional to the probability that the exciton is excited, which is given by

$$P_\mathrm{f}(t) = |a_\mathrm{f}(t)|^2 = \frac{(eE/2\hbar)^2 \, |x_\mathrm{if}|^2}{(eE/2\hbar)^2 \, |x_\mathrm{if}|^2 + (\epsilon/2)^2} \sin^2 \Omega t.$$

The Rabi oscillations in this probability reach an amplitude of one if the excitation is resonant. The amplitude of the oscillations as a function of the detuning ϵ is given by a Lorentz curve. On resonance ($\epsilon = 0$), we have $\Omega = eE|x_{if}|/2\hbar$, i.e. the frequency of the Rabi oscillations depends on the power of the incident radiation (via the electric field strength E).

If the excitation is switched off after half an oscillation period, the system is in the state $|X\rangle$ with certainty. Pulses of this duration are therefore called π-pulses because they flip the qubit state. If the pulse duration is only a quarter of the oscillation period, the system is in a coherent superposition of $|0\rangle$ and $|X\rangle$ ($\pi/2$-pulse).

It is important for this experiment that the pulses have a duration that is small compared to the decoherence time $\tau_\varphi > 500\,\mathrm{ps}$, the recombination time $\tau_\mathrm{rec} \sim 1\,\mathrm{ns}$ and the tunneling rate given by $\tau_\mathrm{tunnel} \sim 10\,\mathrm{ps}$ in this experiment. The pulse duration used in the experiment was therefore only 1 ps. The repetition rate $1/t_\mathrm{r}$ was 82 MHz. In the experiment, the population oscillations of the state $|X\rangle$ were not measured as a function of t_p, but as a function of the applied laser power. Figure 22.16 shows the first oscillation period. The decay of the oscillation amplitude is again due to undesired decoherence mechanisms.

22.4.3 Quantum dot spin-qubits

If we envisage to use the Zeeman-split spin states of a single electron in a quantum dot as a qubit (called a spin-qubit), a hamiltonian allowing us to perform all conceivable unitary single-qubit operations is the Zeeman hamiltonian (Loss and DiVincenzo, 1998)

$$\hat{H} = -\frac{1}{2} g^\star(t) \mu_\mathrm{B} \sigma \mathbf{B}(t) = -\frac{1}{2} g^\star(t) \mu_\mathrm{B} \begin{pmatrix} B_z & B_x - iB_y \\ B_x + iB_y & -B_z \end{pmatrix}. \tag{22.18}$$

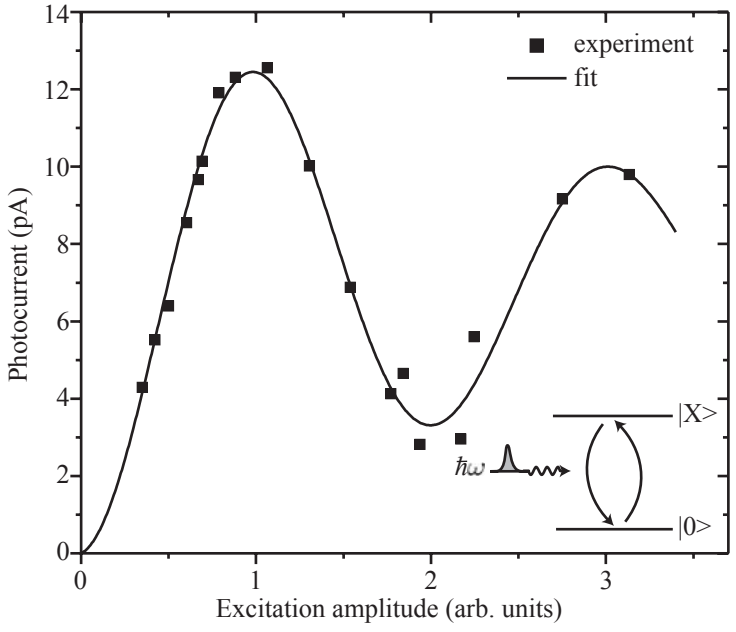


Fig. 22.16 Rabi oscillations of an exciton in a self-assembled InGaAs quantum dot. (Zrenner *et al.*, 2002. Reprinted by permission from Macmillan Publishers Ltd, copyright 2002.)

Here we have introduced a time-dependent g-factor (meaning the effective g-factor in a semiconductor) and a time-dependent magnetic field, in order to indicate what the parameters are that can possibly be controlled externally. While it is obvious that the magnetic field can be an externally controlled parameter in an experiment on a qubit, it is less obvious how the g-factor can be changed in time. A proposal as to how this can be achieved is depicted in Fig. 22.17. The basic idea here is that the electron is confined in a carefully designed layered material system composed of layers with different effective g-factor. The g-factor experienced by the electron can be changed in time if suitable gate voltages allow us to shift the confined electronic wave function from one layer to the other in a time-controlled manner.

We now aim to deduce an equation of motion for the qubit state expressed in terms of the polarization vector $\mathbf{P}$. The equation of motion

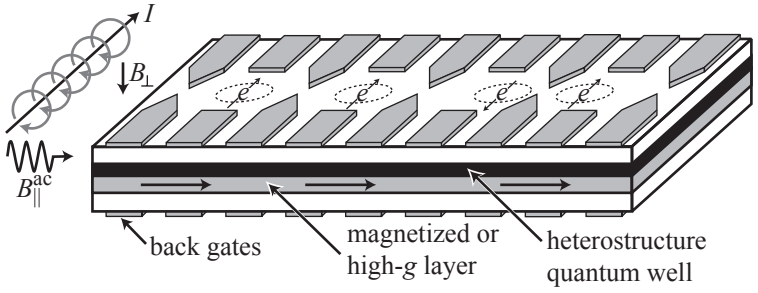


Fig. 22.17 Prototype of a quantum computer as proposed by theorists. (Burkard *et al.*, 2000, see also Cerletti *et al.*, 2005. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

for the density matrix is the von Neumann equation

$$i\hbar\partial_t\hat{\rho} = [\hat{H}, \hat{\rho}] = \hat{H}\hat{\rho} - \hat{\rho}\hat{H},$$

which is equivalent to the Schrödinger equation for wave functions. For a two-level system like a qubit, the hamiltonian can often be written in the form

$$\hat{H} = \begin{pmatrix} \Delta/2 & t \\ t^\star & -\Delta/2 \end{pmatrix},$$

if the energy zero is chosen appropriately. Comparing with eq. (22.18) we find equivalence with the Zeeman hamiltonian, if we define an effective magnetic field

$$B_x = (t + t^\star)/g^\star\mu_\mathrm{B}, \quad B_y = (t^\star - t)/ig^\star\mu_\mathrm{B}, \quad B_z = \Delta/g^\star\mu_\mathrm{B}.$$

If we explicitly work out the commutator on the right-hand side of the von Neumann equation and introduce the polarization vector, we find

$$\begin{pmatrix} i(B_xP_y - B_yP_x) & -(B_x - iB_y)P_z + (P_x - iP_y)B_z \\ (B_x + iB_y)P_z - B_z(P_x + iP_y) & -i(B_xP_y - B_yP_x) \end{pmatrix},$$

where we have omitted the prefactor $-g^\star\mu_\mathrm{B}/2$. Inserting the polarization vector also on the left-hand side of the von Neumann equation, we obtain the equations for the components of the polarization vector

$$\begin{aligned} \hbar\partial_t P_x &= -g^\star\mu_\mathrm{B}(B_yP_z - B_zP_y)/2 \\ \hbar\partial_t P_y &= -g^\star\mu_\mathrm{B}(B_zP_x - B_xP_z)/2 \\ \hbar\partial_t P_z &= -g^\star\mu_\mathrm{B}(B_xP_y - B_yP_x)/2. \end{aligned}$$

This can be written in short as

$$\partial_t\mathbf{P}(t) = \frac{g^\star(t)\mu_\mathrm{B}}{2\hbar}\mathbf{P}(t) \times \mathbf{B}(t).$$

This equation is called the *Bloch equation*. It corresponds to the classical equation of motion of a magnetic moment in an external magnetic field. For example, if the field $\mathbf{B}$ and $g^\star$ are constant in time, this equation describes the Larmor precession. In our context, it is the equation of motion for a qubit under the influence of a very general, possibly time-dependent, hamiltonian. Indeed, we could have described the free oscillation of the qubit in terms of Bloch's equation with a time-independent $\mathbf{B}$ and $g^\star$. We could also have described the Rabi oscillations of the excitonic qubit in terms of Bloch's equation, by choosing the appropriate time-dependent effective magnetic field. If we have a real magnetic field with the time dependence

$$\mathbf{B}(t) = \begin{pmatrix} 0 \\ 0 \\ B_0 \end{pmatrix} + \begin{pmatrix} B_1\cos(\omega t) \\ B_1\sin(\omega t) \\ 0 \end{pmatrix},$$

and $g^\star$ is independent of time, i.e., there is a static field in z-direction of strength B_0 and a field rotating in the x-y plane with angular velocity

ω, then Bloch's equation describes the magnetic resonance phenomenon. These few examples show the versatility of Bloch's equation for the description of the time evolution of a single qubit.

Inelastic relaxation and decoherence can be included in an extended version of the Bloch equation in an empirical way by regarding it as a kind of rate equation for $\mathbf{P}$. To this end we introduce relaxation times T_1 and T_2 and write the equation of motion

$$\partial_t \mathbf{P}(t) = \frac{g^\star(t)\mu_\mathrm{B}}{2\hbar}\mathbf{P}(t) \times \mathbf{B}(t) + \begin{pmatrix} -P_x/T_2 \\ -P_y/T_2 \\ -(P_z - P_z^{(0)})/T_1 \end{pmatrix}.$$

In order to get insight into the meaning of the relaxation times and the new parameter $P_z^{(0)}$, we consider the simple example of $\mathbf{B} = (0, 0, B_0)$. The stationary solution ($\partial_t \mathbf{P} = 0$) of the extended Bloch equations is $\mathbf{P} = (0, 0, P_z^{(0)})$. This stationary solution describes the equilibrium state of the qubit. For example, consider the qubit to interact with a heat bath at temperature T. The ratio of the occupation probabilities p_0 and p_1 of the two Zeeman split levels in thermodynamic equilibrium of the qubit with the bath is given by

$$\frac{p_0}{p_1} = e^{-\Delta/k_\mathrm{B}T},$$

and $P_z = 2p_0 - 1 = 1 - 2p_1$. For very low temperatures $k_\mathrm{B}T \ll \Delta$, $p_0 \to 0$ and $p_1 \to 1$. The stationary ground state of the qubit is therefore $\mathbf{P} = (0, 0, -1)$. If we start at time $t = 0$ with the qubit in the state $\mathbf{P} = (1, 0, 0)$ (meaning $p_0 = 1$ and $p_1 = 0$, qubit energy $\Delta/2$) the qubit is not in thermal equilibrium with the bath. According to the extended Bloch equation, the polarization vector will decay according to $\mathbf{P}(t) = (0, 0, 2e^{-t/T_1} - 1)$ into the thermal ground state (meaning $p_0 = 0$ and $p_1 = 1$, qubit energy $-\Delta/2$). Doing this, the qubit loses the energy Δ. The relaxation time T_1 is therefore called the *inelastic relaxation time*.

If the temperature of the thermal bath is $k_\mathrm{B}T \gg \Delta$, $p_0/p_1 \to 1$ and the stationary ground state of the qubit is the completely mixed state $\mathbf{P} = (0, 0, 0)$. If we start at time $t = 0$ again with the qubit in the state $\mathbf{P} = (1, 0, 0)$, the qubit is again not in thermal equilibrium with the bath and the polarization vector will decay according to the extended Bloch equation as $\mathbf{P}(t) = (0, 0, e^{-t/T_1})$ into the thermal ground state.

In order to see the meaning of T_2 we assume a situation where T_1 is very long, i.e., $T_1 \to \infty$. In this case, the z-component of the relaxation term in the extended Bloch equation is zero, i.e., there is no inelastic relaxation. We further assume that $\mathbf{B} = (0, 0, B_0)$ is time-independent and that we start the qubit in the superposition state $\mathbf{P} = (1, 0, 0)$ at $t = 0$. This is the case of the free qubit oscillation. For this direction of the magnetic field, the z-component of $\mathbf{P}$ will be stationary, i.e., $P_z = 0$ for all times. The problem therefore reduces to the determination of

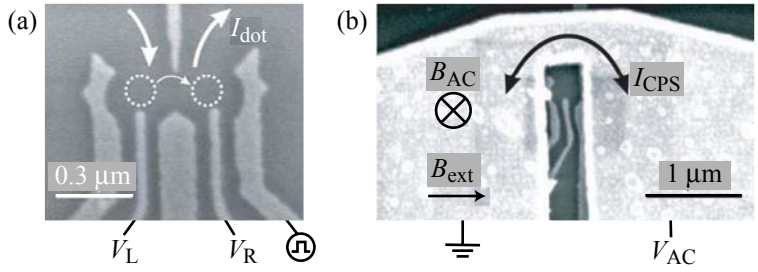


Fig. 22.18 Device designed for the spin-resonance experiment. (a) The double quantum dot defined in a first step on a Ga[Al]As heterostructure with a two-dimensional electron gas by the deposition of suitably shaped metallic top gates. (b) After covering the double quantum dot structure with a thin insulator, the shortcutted end of an RF stripline is deposited above the double dot. (Koppens *et al.*, 2006. Reprinted by permission from Macmillan Publishers Ltd, copyright 2006.)

$P_x(t)$ and $P_y(t)$. The corresponding equations are

$$\begin{aligned} \partial_t P_x &= \omega_0 P_y - P_x/T_2 \\ \partial_t P_y &= -\omega_0 P_x - P_y/T_2, \end{aligned}$$

where $\omega_0 = g^\star \mu_B B_0/2\hbar$. We find the characteristic equation of this system of differential equations with the *Ansatz* $\mathbf{P} = \mathbf{P}_0 e^{i\omega t}$ and get the solutions $\omega_\pm = \pm\omega_0 + i/T_2$. The corresponding eigenvectors are $(1, i)$ for $+$ and $(1, -i)$ for $-$. The general solution can therefore be written as

$$\mathbf{P}(t) = \left[A \begin{pmatrix} 1 \\ i \end{pmatrix} e^{i\omega_0 t} + B \begin{pmatrix} 1 \\ -i \end{pmatrix} e^{-i\omega_0 t} \right] e^{-t/T_2},$$

with coefficients A and B to be determined from the initial condition at $t = 0$. They are found to be $A = 1/2$ and $B = -1/2$ resulting in

$$\mathbf{P}(t) = \begin{pmatrix} \cos\omega_0 t \\ -\sin\omega_0 t \end{pmatrix} e^{-t/T_2}.$$

The qubit rotates in the equatorial plane of the Bloch sphere with angular velocity ω_0 while the length of the polarization vector $|\mathbf{P}|$ decays exponentially with time constant T_2. This decay means an exponential decay of the off-diagonal elements of the density matrix and is therefore called the *decoherence time*. The final state at large times is given by $\mathbf{P} = (0, 0, 0)$, i.e., by the completely mixed state.

Although an electron spin resonance experiment on an electron in a single quantum dot is conceptually straightforward, it has not been realized so far. However, experiments on double quantum dot devices exist in which electron spin resonance was observed in the spin-blockade regime. The corresponding device is shown in Fig. 22.18. The double quantum dot device is realized on the basis of a two-dimensional electron gas in a remotely doped Ga[Al]As heterostructure by depositing suitable metallic top gates [see Fig. 22.18(a)]. The whole structure is then covered with a thin insulator on which a metallic radio-frequency (RF) strip-line is fabricated. The strip line ends with a short above the quantum dot [see Fig. 22.18(b)]. If an RF voltage is applied to the strip-line, an alternating current flows through the short and creates an alternating magnetic field at the position of the double dot.

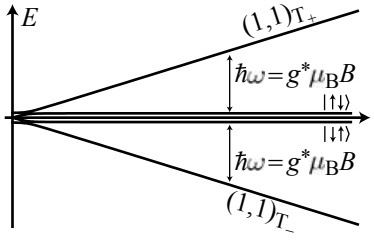


Fig. 22.19 Energy levels of a two-electron double quantum dot under the influence of an external magnetic field B and a small (random) nuclear magnetic field. The latter mixes the $(1,1)_{T_0}$ and the $(1,1)_S$ states into $|\uparrow\downarrow\rangle$ and $|\downarrow\uparrow\rangle$. The spin resonance transitions are indicated by arrows.

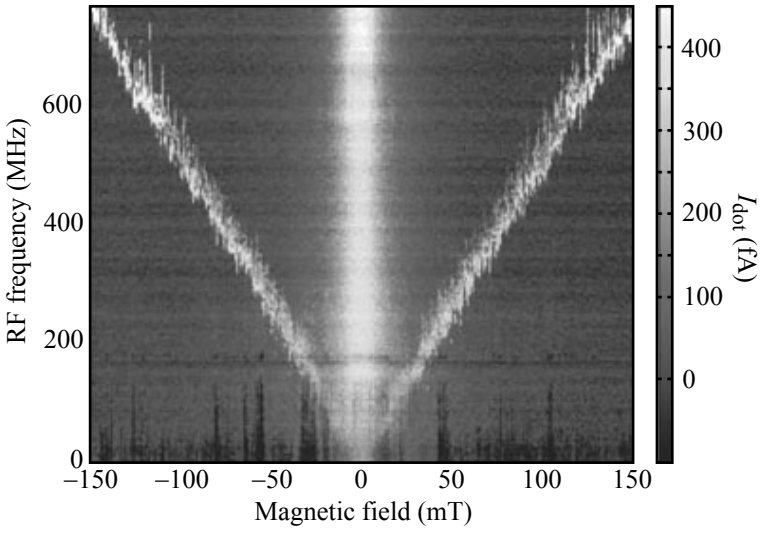


Fig. 22.20 Electron spin resonance measured in the double quantum dot device shown in Fig. 22.18. The system is tuned in the spin-blockade. The resonant microwave field lifts the spin-blockade and a finite current is detected. (Koppens *et al.*, 2006. Reprinted by permission of Macmillan Publishers Ltd, copyright 2006.)

The system is adjusted in a situation with negative detuning and finite static magnetic field parallel to the sample surface [see Fig. 19.8(b)] where the states $(1,1)_{\mathrm{S}}$ and $(1,1)_{\mathrm{T}_0}$ are still degenerate, but $(1,1)_{\mathrm{T}_+}$ and $(1,1)_{\mathrm{T}_-}$ are Zeeman split-off. The energy levels are shown in Fig. 22.19 as a function of magnetic field. In the spin-blockade situation, the system gets stuck either in the $(1,1)_{\mathrm{T}_-}$ or in the $(1,1)_{\mathrm{T}_+}$ state. A magnetic field oscillating with a frequency ω close to resonance (i.e., the Zeeman energy) couples these states with the $(1,1)_{\mathrm{T}_0}$ state. This state is mixed with the $(1,1)_{\mathrm{S}}$ state by a small randomly oriented magnetic field originating from the nuclear spins in the two quantum dots (see discussion on page 418). As a consequence, the spin-blockade is lifted by the application of a resonant RF signal on the strip-line, and a transport current can be detected. The result of the corresponding measurement is shown in Fig. 22.20. The resonance frequency increases linearly with magnetic field, as expected from the resonance condition $\hbar\omega = g^\star \mu_{\mathrm{B}} B$. The frequency-independent peak in the current around zero magnetic field is a result of the singlet–triplet mixing caused by the nuclear field.

The same experimental setup can also be used for a measurement of coherent Rabi oscillations if the alternating magnetic field is pulsed. To this end, the system is prepared in the spin-blockade, where it can be described as a statistical mixture of $(1,1)_{\mathrm{T}_+}$ and $(1,1)_{\mathrm{T}_-}$. The RF magnetic field pulse leads to Rabi oscillations with $(1,1)_{\mathrm{T}_0}$. Depending on the duration of the pulse, the system ends up in any superposition of the two. The measurement is performed by detecting the average current created by a large sequence of RF pulses. The result of such a measurement is shown in Fig. 22.21. As expected, the measured current oscillates as a function of pulse duration, and the oscillation frequency depends on the applied RF power, giving larger frequencies for larger power.

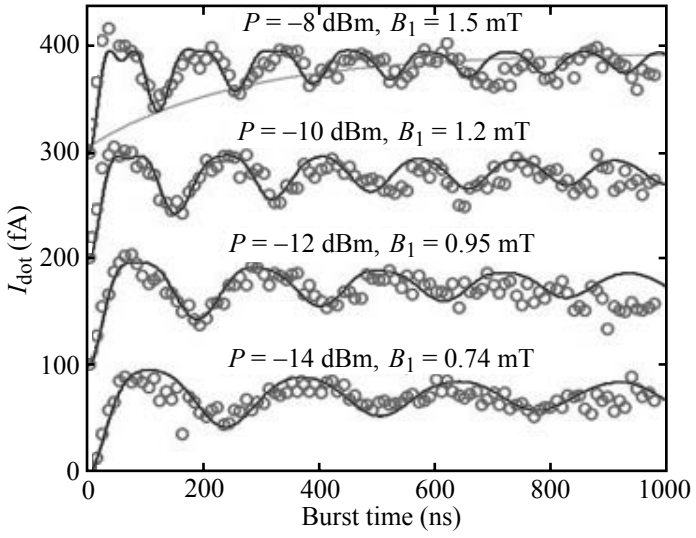


Fig. 22.21 Rabi oscillations measured as a function of RF pulse duration (burst time) and RF power. (Koppens *et al.*, 2006. Reprinted by permission of Macmillan Publishers Ltd, copyright 2006.)

Later experiments (Nowack *et al.*, 2007) have shown that coherent oscillations between Zeeman-split spin states in the spin-blockade situation can also be induced by applying RF voltages to one of the gate electrodes. This phenomenon is called *electrically driven spin resonance* (EDSR). It is believed that the orbital motion of the electron induced by the electric field translates into an effective magnetic field via spin–orbit interaction.

Two electrons in a double quantum dot as a two spin-qubit system

The controlled manipulation of a two-electron double quantum dot allows us to perform a two-qubit operation on the two electron spins called a swap operation (Petta *et al.*, 2005). For explaining the sequence of steps required for the swap operation, we use Fig. 19.8(b). A magnetic field splits the three (1,1) triplet states into $(1,1)_{\mathrm{T}_+}$, $(1,1)_{\mathrm{T}_0}$, and $(1,1)_{\mathrm{T}_-}$. In the first step the system is initialized in the $(0,2)_{\mathrm{S}}$ spin singlet ground state. In the second step, the detuning is quickly swept slightly beyond the crossing of the $(0,2)_{\mathrm{S}}$ and the $(1,1)_{\mathrm{T}_-}$ states in order to avoid an uncontrolled spin flip at the crossing point. In the third step, the detuning between the two dots is further increased adiabatically, i.e., slowly, such that the system evolves into the region where the $(1,1)_{\mathrm{S}}$ state is degenerate with the $(1,1)_{\mathrm{T}_0}$ state. The adiabatic change of δ allows the spins to align along the nuclear magnetic field and to form the ground state $|\uparrow\downarrow\rangle$ which is a superposition of $(1,1)_{\mathrm{S}}$ and $(1,1)_{\mathrm{T}_0}$. Once the system is settled in this state, pulsing the detuning close (but not beyond) the $(1,1)_{\mathrm{T}_-}$ state for a finite time t leaves the system in this superposition, but the two components have energies differing by a finite amount $\Delta(\delta)$. This is the situation in which free oscillations between

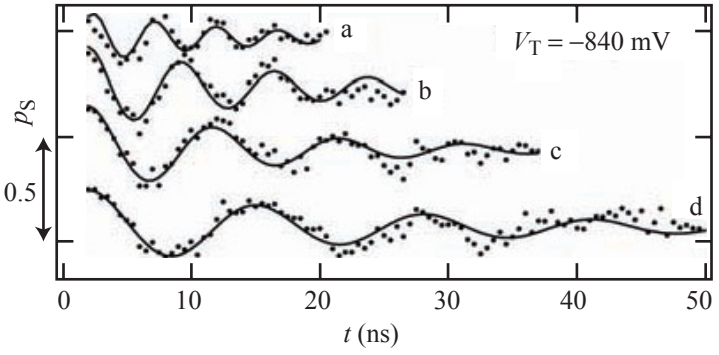


Fig. 22.22 Free oscillations between the the $(1,1)_S$ and the $(1,1)_{T_0}$ state for the energy splitting $\Delta(\delta)$ decreasing from a to d (Petta *et al.*, 2005).

the two superimposed states occur in close resemblance to eq. (22.16). A relative phase $\Delta(\delta)t/\hbar$ develops over time. If the dwell time at this detuning is chosen to be $t_E = \pi\hbar/\Delta(\delta)$ (π-pulse), the spins have evolved from $|\uparrow\downarrow\rangle$ to $|\downarrow\uparrow\rangle$. This is the desired swap operation of the two spins. The spin state can be read out with a charge detector by going through steps one to four in reverse order after this pulse. The detected occupation p_S of $(0,2)_S$ is maximum after a 2π-pulse, while a π-pulse leads to minimum occupation of this state. Therefore p_S is found to oscillate as a function of pulse duration t, as depicted in Fig. 22.22.

Further reading

- Books on classical information and computation: Brillouin 1956; Feynman 1996; Cover and Thomas 1991.
- Paper on statistical mechanics and information: Jaynes 1957.
- Books on quantum information and computation: Williams and Clearwater 1997; Lo *et al.* 1998; Nielsen and Chuang 2000; Vedral 2006; Mermin 2007.
- Lecture notes on quantum information and computation: Berthiaume 1997; Preskill 1998; Ekert *et al.* 2001.
- Reviews of quantum information and computation: Steane 1998.
- Quantum information with spins in quantum dots: Burkard *et al.* 2000; Cerletti *et al.* 2005.
- Einstein–Podolsky–Rosen paradox and Bell inequalities: Bell 1964; Aspect 2002.

Exercises

(22.1) In this exercise you work on a data compression problem. We consider an information source which randomly picks letters from the string *physics is good for you* and forms a stream of these letters without spaces. An example would be emitter→*pyokosuphgo...*→receiver. How many bits are required to save the data stream if

(a) each letter is encoded with four bits,

(b) the letters are encoded with an optimized bit sequence,

(c) the limit of the Shannon entropy could be reached?

Discuss how many qubits would be needed to save the data stream. Write down the letters and the corresponding codes in a table and calculate in all cases the Shannon entropy. Hint: For the decoding process, the receiver must know when a new letter begins. What are the implications for a sequence of numbers without spaces? Emitter and receiver can use certain rules that they agree on beforehand.

(22.2) A density matrix ρ describes the statistical state of a quantum system. The need for a statistical description arises when one considers either an ensemble of systems or one system when its preparation history is uncertain. The expectation value $\langle A\rangle$ of any observable A can be calculated using the density matrix. It is given by

$$\langle A\rangle = \mathrm{Tr}[\rho A].$$

(a) The ensemble is in the state $|\,\psi\rangle$. Give an explicit expression of the density matrix in Dirac notation.

(b) The state $|\,\psi\rangle$ can be expanded in eigenstates of the observable A, $|\,\psi\rangle = \sum_n a_n\,|\,n\rangle$, where $\langle n|m\rangle = \delta_{nm}$. Give an alternative expression for ρ using the eigenstates of A.

(c) The probabilities $p_n = |a_n|^2$ can be used to specify the von Neumann entropy S,

$$S = -\sum_n p_n \log_2 p_n.$$

This relation is in close analogy to the Shannon entropy in information theory. Express S in terms of the density matrix.

(d) What is the von Neumann entropy of a pure state? How does a unitary transformation (the time evolution of a quantum system) change the entropy? How do you interpret the result in the light of reversibility?

Fourier transform and Fourier series

A

A.1 Fourier series of lattice periodic functions

Let $u(\mathbf{r}) = u(\mathbf{r} + \mathbf{R})$ be a lattice periodic function. Its expansion in a Fourier series is

$$u(\mathbf{r}) = \sum_{\mathbf{K}} c_{\mathbf{K}} e^{i\mathbf{K}\mathbf{r}}$$

with

$$c_{\mathbf{K}} = \frac{1}{V_0} \int_{\mathrm{UC}} d^3r \;\; u(\mathbf{r}) e^{-i\mathbf{K}\mathbf{r}},$$

where the space integration has to be taken over the unit cell (UC) with volume V_0, and $\mathbf{K}$ is a vector of the reciprocal lattice.

A.2 Fourier transform

The Fourier transform of a function is given by

$$U(\mathbf{r}) = \int \frac{d^3k}{(2\pi)^3} \, U(\mathbf{k}) e^{i\mathbf{k}\mathbf{r}}$$

with the inverse transform

$$U(\mathbf{k}) = \int d^3r \, U(\mathbf{r}) e^{-i\mathbf{k}\mathbf{r}}.$$

A.3 Fourier transform in two dimensions

In two dimensions the Fourier transform of a function is given by

$$U(\mathbf{r}) = \int \frac{d^2k}{(2\pi)^2} \, U(\mathbf{k}) e^{i\mathbf{k}\mathbf{r}}$$

with the inverse transform

$$U(\mathbf{k}) = \int d^2r \;\; U(\mathbf{r}) e^{-i\mathbf{k}\mathbf{r}}.$$

If the function $U(\mathbf{r})$ possesses radial symmetry, i.e. it depends only on $r = |\mathbf{r}|$, then

$$U(k) = \int_0^\infty dr \;\; rU(r) \int_0^{2\pi} d\varphi \;\; e^{-ikr\cos\varphi} = 2\pi \int_0^\infty dr \;\; rJ_0(kr)U(r),$$

and correspondingly for the inverse transform

$$\begin{aligned} U(r) &= \frac{1}{(2\pi)^2} \int_0^\infty dk \;\; kU(k) \int_0^{2\pi} d\varphi \;\; e^{ikr\cos\varphi} \\ &= \frac{1}{2\pi} \int_0^\infty dk \;\; kJ_0(kr)U(k). \end{aligned}$$

We talk about the Fourier–Bessel expansion, because the $J_0(kr)$ are Bessel functions.

Fourier transform of the Coulomb potentials. The two-dimensional Fourier transform of the Coulomb potential

$$U(\mathbf{r}, z) = \frac{1}{\sqrt{r^2 + z^2}}$$

is given by

$$U(k) = 2\pi \int_0^\infty dr \; rJ_0(kr) \frac{1}{\sqrt{r^2 + z^2}} = \frac{2\pi}{k} e^{-kz}.$$

Extended Green's theorem and Green's function

B

B.1 Derivation of an extended version of Green's theorem

In order to solve eq. (7.1) formally, we need an extended version of Green's theorem which we will derive in the following (Smythe, 1939): let ϕ, ϵ, and ψ be any scalar functions of the space coordinates $\mathbf{r}$. Then

$$\nabla \left[\phi\epsilon\nabla\psi\right] = \phi\nabla \left[\epsilon\nabla\psi\right] + \nabla\phi\epsilon\nabla\psi.$$

Application of Gauss's integral theorem leads to

$$\oint_S ds \left[\phi\epsilon\nabla\psi\right]\mathbf{n} = \int_V dV \left\{\phi\nabla \left[\epsilon\nabla\psi\right] + \nabla\phi\epsilon\nabla\psi\right\}.$$

Here $\mathbf{n}$ is the outer normal of the surface element ds. Interchanging ψ and ϕ, we obtain the relation

$$\oint_S ds \left[\psi\epsilon\nabla\phi\right]\mathbf{n} = \int_V dV \left\{\psi\nabla \left[\epsilon\nabla\phi\right] + \nabla\psi\epsilon\nabla\phi\right\}.$$

The difference of the two latter equations is

$$\oint_S ds \left[\psi\epsilon\nabla\phi - \phi\epsilon\nabla\psi\right]\mathbf{n} = \int_V dV \left\{\psi\nabla \left[\epsilon\nabla\phi\right] - \phi\nabla \left[\epsilon\nabla\psi\right]\right\}. \qquad \text{(B.1)}$$

This is the desired extended version of Green's integral theorem.

B.2 Proof of the symmetry of Green's functions

Green's function has the property

$$G(\mathbf{r}_1, \mathbf{r}_2) = G(\mathbf{r}_2, \mathbf{r}_1).$$

In order to prove this (Meetz and Engl, 1980) we replace in eq. (B.1) ψ by $G(\mathbf{r}_1, \mathbf{r}_2)$, ϵ by $\epsilon(\mathbf{r})\epsilon_0$, and ϕ by $G(\mathbf{r}_1, \mathbf{r}_3)$. We obtain

$$\begin{aligned}\oint_S ds_1 \left[G(\mathbf{r}_1, \mathbf{r}_2)\epsilon(\mathbf{r}_1)\epsilon_0\nabla_1 G(\mathbf{r}_1, \mathbf{r}_3) - G(\mathbf{r}_1, \mathbf{r}_3)\epsilon(\mathbf{r}_1)\epsilon_0\nabla_1 G(\mathbf{r}_1, \mathbf{r}_2)\right]\mathbf{n} \\ = \int_V dV_1 \left\{G(\mathbf{r}_1, \mathbf{r}_2)\nabla_1 \left[\epsilon(\mathbf{r}_1)\epsilon_0\nabla G(\mathbf{r}_1, \mathbf{r}_3)\right]\right. \\ \left. - G(\mathbf{r}_1, \mathbf{r}_3)\nabla_1 \left[\epsilon(\mathbf{r}_1)\epsilon_0\nabla_1 G(\mathbf{r}_1, \mathbf{r}_2)\right]\right\}.\end{aligned}$$

As a result of the boundary conditions for Green's function, the surface integral on the left-hand side gives zero. The right-hand side simplifies if we consider the definition of Green's function. We obtain

$$0 = \int_V dV_1 \left\{ G(\mathbf{r}_1, \mathbf{r}_2)\delta(\mathbf{r}_1 - \mathbf{r}_3) - G(\mathbf{r}_1, \mathbf{r}_3)\delta(\mathbf{r}_1 - \mathbf{r}_2) \right\}$$
$$= G(\mathbf{r}_3, \mathbf{r}_2) - G(\mathbf{r}_2, \mathbf{r}_3).$$

This completes the proof of the symmetry of Green's function with respect to interchanging the two arguments.

The delta-function

C

Dirac's delta function is a generalization of the Kronecker-symbol

$$\delta_{nm} = \begin{cases} 0 & \text{for } m \neq n \\ 1 & \text{for } m = n \end{cases} .$$

Formally,

$$\delta(x-y) = \begin{cases} 0 & \text{for } x \neq y \\ \infty & \text{for } x = y \end{cases} .$$

However, the integral of the delta function gives

$$\int_{-\infty}^{+\infty} \delta(x-y)dx = 1.$$

The delta function can be represented as the limiting case of integral operators. For example,

$$\begin{aligned} \delta(x-y) &= \frac{1}{\pi} \lim_{k\to\infty} \frac{\sin[k(x-y)]}{x-y} \\ &= \frac{1}{\pi} \lim_{\epsilon\to 0^+} \frac{\epsilon}{(x-y)^2+\epsilon^2}. \end{aligned}$$

Furthermore,

$$\lim_{\epsilon\to 0^+} \frac{1}{x-y\pm i\epsilon} = \mathrm{P}\frac{1}{x-y} \mp i\pi\delta(x-y).$$

Fundamental properties of the delta function are

$$\begin{aligned} \delta(x) &= \delta(-x) \\ \delta(ax) &= \frac{1}{|a|}\delta(x) \\ \delta(f(x)) &= \sum_n \frac{1}{|f(x_n)|}\delta(x-x_n), \text{ where } f(x_n)=0 \\ f(x)\delta(x-a) &= f(a)\delta(x-a) \\ \int \delta(x-y)\delta(y-a)dy &= \delta(x-a) \\ \delta(x) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{ikx}dk. \end{aligned}$$

In cases where derivatives of the delta functions appear, the following

rules can be applied:

$$\begin{aligned}\int_{-\infty}^{+\infty} \delta^{(n)}(x) f(x) &= (-1)^n f^{(n)}(0) \\ \int_{-\infty}^{+\infty} \delta'(x) f(x) &= -f'(0) \\ \delta'(x) &= \frac{i}{2\pi} \int_{-\infty}^{+\infty} k e^{ikx} dk.\end{aligned}$$

References

Abrahams, E., Anderson, P.W., Licardello, D.C., and Ramakrishnan, T.V. (1979). *Phys. Rev. Lett.*, **42**, 673.

Abramowitz, M. and Stegun, I.A. (1984). *Pocketbook of Mathematical Functions*. Harry Deutsch, Frankfurt/Main.

Adachi, S. (1985). *J. Appl. Phys.*, **58**, R1.

Aeberhard, U., Wakabayashi, K., and Sigrist, M. (2005). *Phys. Rev. B*, **72**, 075328.

Agam, O., Aleiner, I., and Larkin, A. (2000). *Phys. Rev. Lett.*, **85**, 3153.

Aharonov, Y. and Bohm, D. (1959). *Phys. Rev.*, **115**, 485.

Aharonov, Y. and Casher, A. (1984). *Phys. Rev. Lett.*, **53**, 319.

Aleiner, I.L., Brouwer, P.W., and Glazman, L.I. (2002). *Phys. Rep.*, **358**, 309.

Alhassid, Y. (2000). *Rev. Mod. Phys.*, **72**, 895.

Altshuler, B.L. (1985). *JETP Lett.*, **41**, 648.

Altshuler, B.L. and Aronov, A.G. (1985). Electron–electron interactions in disordered conductors. In *Electron–electron interactions in disordered systems*, pp. 1–153. Elsevier Science Publishers, Amsterdam.

Altshuler, B.L., Aronov, A.G., and Spivak, B.Z. (1981). *JETP Lett.*, **33**, 101.

Anandan, J. (1982). *Phys. Rev. Lett.*, **48**, 1660.

Ando, T. (1982). Self-consistent results for a GaAs/Al_xGa_{1-x}As heterojunction. i. Subband structure and light scattering spectra. *J. Phys. Soc. Japan*, **51**, 3893.

Ando, T. (2005). Theory of electronic states and transport in carbon nanotubes. *J. Phys. Soc. Jap.*, **74**, 777.

Ando, T., Arakawa, Y., Furuya, K., Komiyama, S., and Nakashima, H. (1998). *Mesoscopic Physics and Electronics*. Nanoscience and Nanotechnology. Springer, Berlin.

Ando, T., Fowler, A.B., and Stern, F. (1982). Electronic properties of two-dimensional systems. *Rev. Mod. Phys.*, **54**, 437.

Aoki, H. (1977). *J. Phys. C*, **10**, 2583.

Aoki, H. (1979). *J. Phys. C*, **12**, 633.

Aoki, H. and Ando, T. (1981). *Solid State Commun.*, **38**, 1079.

Arndt, M., Nairz, O., Vos-Andreae, J., Keller, C., van der Zouw, G., and Zeilinger, A. (1999). *Nature*, **401**, 680.

Ashcroft, N.W. and Mermin, N.D. (1987). *Solid State Physics* (int. edn). Saunders College, Philadelphia.

Aspect, A. (2002). Bell's theorem: The naive view of an experimentalist. In *Quantum [Un]speakables – From Bell to Quantum Information* (ed. A. Bertlmann and A. Zeilinger). Springer, Berlin.

Averin, D.V., Korotkov, A.N., and Likharev, K.K. (1991). *Phys. Rev. B*, **44**, 6199.

Averin, D.V. and Nazarov, Y.V. (1990). *Phys. Rev. Lett.*, **65**, 2446.

Averin, D.V. and Sukhorukov, E.V. (2005). *Phys. Rev. Lett.*, **95**, 126803.

Avinun–Kalish, M., Heiblum, M., Zarchin, O., Mahalu, D., and Umansky, V. (2005). *Nature*, **436**, 529.

Bagrets, D.A. and Nazarov, Y.V. (2003). *Phys. Rev. B*, **67**, 085316.

Balkanski, M. and Wallis, R.F. (2000). *Semiconductor Physics and Applications*. Oxford University Press, Oxford.

Bardeen, J. (1961). *Phys. Rev. Lett.*, **6**, 57.

Barenco, A., Bennett, C.H., Cleve, R., DiVincenzo, D.P., Margolus, N., Shor, P., Sleator, T., Smolin, J.A., and Weinfurter, H. (1995). *Phys. Rev. A*, **52**, 3457.

Beenakker, C.W.J. (1990). *Phys. Rev. Lett.*, **64**, 216.

Beenakker, C.W.J. (1991). *Phys. Rev. B*, **44**, 1646.

Beenakker, C.W.J. (1997). *Rev. Mod. Phys.*, **69**, 731.

Beenakker, C.W.J. and Buttiker, M. (1992). *Phys. Rev. B*, **46**, 1889.

Beenakker, C.W.J. and Schonenberger, C. (2003). *Physics Today*, **May issue**, 37.

Beenakker, C.W.J. and van Houten, H. (1988*a*). *Phys. Rev. B*, **37**, 6544.

Beenakker, C.W.J. and van Houten, H. (1988*b*). *Phys. Rev. B*, **38**, 3232.

Beenakker, C.W.J. and van Houten, H. (1991). Quantum transport in semiconductor nanostructures. Volume 44 of *Solid State Physics*. Academic Press, Inc.

Bell, J. (1964). *Physics*, **1**, 195.

Belzig, W. (2005). *Physik Journal*, **4**, 75.

Bennett, C.H. (1979). *IBM J. Res. Dev.*, **6**, 525.

Bennett, C.H. (1982). *Int. J. Theor. Phys.*, **21**, 905.

Bergmann, G. (1982). *Solid State Commun.*, **42**, 815.

Bergmann, G. (1983). *Phys. Rev. B*, **28**, 2914.

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

Bergsten, T., Kobayashi, T., Sekine, Y., and Nitta, J. (2006). *Phys. Rev. Lett.*, **97**, 196803.

Berry, M.V. (1984). *Proc. R. Soc. London A*, **392**, 45.

Berry, M.V. (1988). *Scientific American*, **259**, 26.

Berthiaume, Andre (1997). Quantum computation. In *Complexity Theory Retrospective II* (ed. L. Hemaspaandra and A. Selman), pp. 23–51. Springer-Verlag, Berlin Germany.

Birk, H., Jong, M.J.M. De, and Schonenberger, C. (1995). *Phys. Rev. Lett.*, **75**, 1610.

Bitter, T. and Dubbers, D. (1987). *Phys. Rev. Lett.*, **87**, 251.

Bjork, M.T., Ohlsson, B.J., Sass, T., Persson, A.I., Thelander, C., Magnusson, M.H., Deppert, K., Wallenberg, L.R., and Samuelson, L. (2002). *Nano Lett.*, **2**, 87.

Blanter, Ya. and Buttiker, M. (2000). *Phys. Rep.*, **336**, 1.

Bonet, E., Deshmukh, M.M., and Ralph, D.C. (2002). *Phys. Rev. B*, **65**, 045317.

Brandes, T., Hausler, W., Jauregui, K., Kramer, B., and Weinmann, D. (1993). *Physica B*, **189**, 16.

Breit, G. and Wigner, E. (1936). *Phys. Rev.*, **49**, 519.

Bremme, L., Ihn, T., and Ensslin, K. (1999). *Phys. Rev. B*, **59**, 7305.

Briggs, A., Guldner, Y., Vieren, J. P., Voos, M., Hirtz, J. P., and Razeghi, M. (1983). *Phys. Rev. B*, **27**, 6549.

Brillouin, L. (1956). *Science and Information Theory.* Academic Press, New York.

Brosig, S., Ensslin, K., Warburton, R. J., Nguyen, C., Brar, B., Thomas, M., and Kroemer, H. (1999). *Phys. Rev. B*, **60**, R13989.

Bryant, G.W. (1987). *Phys. Rev. Lett.*, **59**, 1140.

Buks, E., Schuster, R., Heiblum, M., Mahalu, D., and Umansky, V. (1998). *Nature*, **391**, 871.

Burkard, G., Engel, H.-A., and Loss, D. (2000). *Fortschr. Phys.*, **48**, 965.

Burt, M.G. (1994). *Appl. Phys. Lett.*, **65**, 717.

Buttiker, M. (1986). *Phys. Rev. Lett.*, **57**, 1761.

Buttiker, M. (1990). *Phys. Rev. B*, **41**, 7906.

Bychkov, Y.A. and Rashba, E.I. (1984*a*). *J. Phys. C: Solid State Phys.*, **17**, 6039.

Bychkov, Y.A. and Rashba, E.I. (1984*b*). *JETP Lett.*, **39**, 78.

Cahay, M., McLennan, M., and Datta, S. (1988). *Phys. Rev. B*, **37**, 10125.

Campbell, N. (1909*a*). *Proc Cambr. Phil. Soc.*, **15**, 117.

Campbell, N. (1909*b*). *Proc Cambr. Phil. Soc.*, **15**, 300.

Cerdeira, F., Fjeldly, T.A., and Cardona, M. (1973). *Phys. Rev. B*, **8**, 4734.

Cerletti, V., Coish, W.A., Gywat, O., and Loss, D. (2005). *Nanotechnology*, **16**, R27.

Chakraborty, T. (1999). *Quantum dots: a survey of the properties of artificial atoms*. Elsevier, Amsterdam.

Chakraborty, T. and Pietilainen, P. (1995). *The quantum Hall effects: integral and fractional* (2nd edn). Springer, Berlin.

Chakravarty, S. and Schmid, A. (1986). *Phys. Rep.*, **140**, 193.

Chan, I.H., Fallahi, P., Westervelt, R.M., Maranowski, K.D., and Gossard, A.C. (2003). *Physica E*, **17**, 584.

Chiang, T.-C., Knapp, J.A., Aono, M., and Eastman, D. (1980). *Phys. Rev. B*, **21**, 3513.

Chklovskii, D.B., Shklovskii, B.I., and Glazman, L.I. (1992). *Phys. Rev. B*, **46**, 4026.

Christen, T. and Buttiker, M. (1996). *Phys. Rev. Lett.*, **77**, 143.

Chu, C.S. and Sorbello, R.S. (1989). *Phys. Rev. B*, **40**, 5941.

Cimmino, A., Opat, G.I., Klein, A.G., Kaiser, H., Werner, S.A., Arif, M., and Clothier, R. (1989). *Phys. Rev. Lett.*, **63**, 380.

Ciorga, M., Sachrajda, A.S., Hawrylak, P., Gould, C., Zawadzki, P., Jullian, S., Feng, Y., and Wasilewski, Z. (2000). *Phys. Rev. B*, **61**, R16315.

Clausius, R. (1867). *Abhandlungen uber die mechanische Warmetheorie*. Vieweg und Sohn, Braunschweig.

Clerk, A.A., Waintal, X., and Brouwer, P.W. (2001). *Phys. Rev. Lett.*, **86**, 4636.

Cohen, M.L. and Chelikowski, J. (1989). *Electronic structure and optical properties of semiconductors* (2nd edn). Springer, Berlin–New York.

Coleridge, P.T., Stoner, R., and Fletcher, R. (1989). *Phys. Rev. B*, **39**, 1120.

Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory*. John Wiley & Sons, Inc., New York.

Crommie, M.F., Lutz, C.P., and Eigler, D.M. (1993). *Nature*, **363**, 524.

Darwin, C.G. (1930). *Prog. Cambridge Philos. Soc.*, **27**, 86.

Das, B., Miller, D.C., Datta, S., Reifenberger, R., Hong, W.P., Bhattacharya, P.K., Singh, J., and Jaffe, M. (1989). *Phys. Rev. B*, **39**, 1411.

Datta, S. (1997). *Electronic Transport in Mesoscopic Systems*. Cambridge University Press, Cambridge.

Davies, J.H. (1998). *The Physics of Low Dimensional Semiconductors*. Cambridge University Press, London.

Davies, J.H., Hyldgaard, P., Hershfield, S., and Wilkins, J.W. (1992). *Phys. Rev. B*, **46**, 9620.

Davies, J.H., Larkin, I.A., and Sukhorukov, E.V. (1995). *J. Appl. Phys.*, **77**, 4504.

De Jong, M.J.M. and Beenakker, C.W.J. (1997). Shot noise in mesoscopic systems. In *Mesoscopic Electron Transport*, Volume 345 of *Nato ASI Series E: Applied Sciences*, pp. 225–258. Kluwer Academic Publishers, Dordrecht.

de Picciotto, R., Reznikov, M., Heiblum, M., Umansky, V., Bunin, G., and Mahalu, D. (1997). *Nature*, **389**, 162.

de Picciotto, R., Stormer, H.L., Pfeiffer, L.N., Baldwin, K.W., and West, K.W. (2001). *Nature*, **411**, 51.

Dingle, R.B. (1952). *Proc. R. Soc. London, Ser. A*, **211**, 517.

Dingle, R. and Störmer, H.L. (1978). *Appl. Phys. Lett.*, **33**, 665.

DiVincenzo, D.P. (1995). *Phys. Rev. A*, **51**, 1015.

Dolan, G.J. and Osheroff, D.D. (1979). *Phys. Rev. Lett.*, **43**, 721.

Dresselhaus, G. (1955). *Phys. Rev.*, **100**, 580.

Drude, P. (1900*a*). *Ann. Physik*, **1**, 566.

Drude, P. (1900*b*). *Ann. Physik*, **3**, 369.

Duncan, D.S., Livermore, C., Westervelt, R.M., Maranowski, K.D., and Gossard, A.C. (1999). *Appl. Phys. Lett.*, **74**, 1045.

D'yakonov, M.I. and Perel, V.I. (1971). *Sov. Phys. JETP*, **33**, 1053.

D'yakonov, M.I. and Perel, V.I. (1972). *Sov. Phys. Solid State*, **13**, 3023.

Efros, A.L. (1989). *Solid State Comm.*, **70**, 253.

Efros, A.L., Pikus, F.G., and Burnett, V.G. (1993). *Phys. Rev. B*, **47**, 2233.

Ehrenberg, W. and Siday, R.E. (1949). *Proc. Phys. Soc.*, **B62**, 8–21.

Ekert, A., Hayden, P., and Inamori, H. (2001). Basic concepts in quantum computation. In *Coherent atomic matter waves*, Les Houches Summer School, Chapter 10, pp. 661. Springer, Berlin. Also available at arXiv:quant-ph/0011013.

Ellenberger, C., Ihn, T., Yannouleas, C., Landman, U., Ensslin, K., Driscoll, D.C., and Gossard, A.C. (2006). *Phys. Rev. Lett.*, **96**, 126806.

Elliott, R.J. (1954). *Phys. Rev.*, **96**, 266.

Elzerman, J.M., Hanson, R., van Beveren, L.H. Willems, Witkamp, B., Vandersypen, L.M.V., and Kouwenhoven, L.P. (2004). *Nature*, **430**, 431.

Enderlein, R. and Schenk, A. (1992). *Grundlagen der Halbleiterphysik*. Akademie Verlag, Berlin.

Engels, G., Lange, J., Schapers, Th., and Luth, H. (1997). *Phys. Rev. B*, **55**, R1958.

Fabian, J. and Sarma, S. Das (1999). *J. Vac. Sci. Technol. B*, **17**, 1708.

Faist, J., Capasso, F., Sirtori, C., West, K.W., and Pfeiffer, L.N. (1997). *Nature*, **390**, 589.

Fano, U. (1935). *Nuovo Cimento*, **12**, 156.

Fano, U. (1961). *Phys. Rev.*, **124**, 1866.

Ferry, D. (1998). *Transport in Nanostructures*. Cambridge University Press, Cambridge.

Fertig, H.A. and Halperin, B.I. (1987). *Phys. Rev. B*, **36**, 7969.

Feshbach, H., Peaslee, D.C., and Weisskopf, V.F. (1947). *Phys. Rev.*, **71**, 145.

Feynman, R.P. (1985). *Optics News*, **11**, 11.

Feynman, R.P. (1992). *Journal of Micromechanical Systems*, **1**, 60.

Feynman, R.P. (1996). *Feynman lectures on computation*. Perseus Publishing, Cambridge, Massachusetts.

Feynman, R.P., Leighton, R.B., and Sands, M. (2006). *The Feynman Lectures on Physics* (Definitive edn). Pearson Addison-Wesley, San Francisco.

Field, M., Smith, C. G., Pepper, M., Ritchie, D.A., Frost, J.E.F., Jones, G.A.C., and Hasko, D.G. (1993). *Phys. Rev. Lett.*, **70**, 1311.

Fock, V. (1928). *Z. Physik*, **47**, 446.

Folk, J.A., Potok, R.M., Marcus, C.M., and Umansky, V. (2003). *Science*, **299**, 679.

Ford, C.J.B., Thornton, T.J., Newbury, R., Pepper, M., Ahmed, H., Davies, G. J., and Andrews, D. (1988). *Superlatt. Microstruct.*, **4**, 541.

Fowler, A.B., Fang, F.F., Howard, W.E., and Stiles, P.J. (1966). *Phys. Rev. Lett.*, **16**, 901.

Foxman, E.B., McEuen, P.L., Meirav, U., Wingreen, N.S., Meir, Y., Belk, P.A., Belk, N.R., Kastner, M.A., and Wind, S.J. (1993). *Phys. Rev. B*, **47**, 10020.

Franceschi, S. De, Sasaki, S., Elzerman, J.M., van der Wiel, W.G., Tarucha, S., and Kouwenhoven, L.P. (2001). *Phys. Rev. Lett.*, **86**, 878.

Fredkin, E. and Toffoli, T. (1982). *Int. J. Theor. Phys.*, **21**, 219.

Frustaglia, D. and Richter, K. (2004). *Phys. Rev. B*, **69**, 235310.

Fuhrer, A., Ihn, T., Ensslin, K., Wegscheider, W., and Bichler, M. (2003). *Phys. Rev. Lett.*, **91**, 206802.

Fuhrer, A., Luscher, S., Ihn, T., Heinzel, T., Ensslin, K., Wegscheider, W., and Bichler, M. (2001). *Nature*, **413**, 822.

Fujisawa, T., Hayashi, T., Cheong, H.D., Jeong, Y.H., and Hirayama, Y. (2004). *Physica E*, **21**, 1046.

Ganichev, S.D., Belkov, V. V., Golub, L. E., Ivchenko, E. L., Schneider, P., Giglberger, S., Eroms, J., Boeck, J. De, Borghs, G., Wegscheider, W., Weiss, D., and Prettl, W. (2004). *Phys. Rev. Lett.*, **92**, 256601.

Gefen, Y. (2002). In *Strongly Correlated Fermions and Bosons in Low-Dimensional Disordered Systems*, Dordrecht, pp. 13. Kluwer Academic Publishers.

Geller, M.R. and Vignale, G. (1995). *Physica B*, **212**, 283.

Gerhardts, R.R. (1975). *Z. Phys. B*, **21**, 275 and 285.

Gerhardts, R.R. (1976). *Surf. Sci.*, **58**, 227.

Giaever, I. and Zeller, H.R. (1968). *Phys. Rev. Lett.*, **20**, 1504.

Giuliani, G.F. and Vignale, G.F. (2005). *Quantum Theory of the Electron Liquid*. Cambridge University Press, New York.

Glazman, L.I. and Larkin, I.A. (1991). *Semicond. Sci. Technol.*, **6**, 32.

Glazman, L.I. and Raikh, M.E. (1988). *JETP Lett.*, **47**, 452.

Glew, R.W., Poole, D.A., and Pepper, M. (1981). *J. Phys. C*, **14**, L395.

Gold, A. and Dolgopolov, V.T. (1986). *Phys. Rev. B*, **33**, 1076.

Goldhaber-Gordon, D., Shtrikman, H., Mahalu, D., Abusch-Magder, D., Meirav, U., and Kastner, M.A. (1998). *Nature*, **391**, 156.

Gores, J., Goldhaber-Gordon, D., Heemeyer, S., Kastner, M.A., Shtrikman, H., Mahalu, D., and Meirav, U. (2000). *Phys. Rev. B*, **62**, 2188.

Gorter, C.J. (1951). *Physica*, **17**, 777.

Grabert, H. and Devoret, M.H. (1992). *Single charge tunneling: Coulomb blockade phenomena in nanostructures*, Volume 294 of *NATO ASI series. Series B, physics*. Plenum Press, New York.

Grbic, B., Leturcq, R., Ihn, T., Ensslin, K., Reuter, D., and Wieck, A. D. (2007). *Phys. Rev. Lett.*, **99**, 176803.

Grundler, D. (2000). *Phys. Rev. Lett.*, **84**, 6074.

Gunnarsson, O. and Lundqvist, B.I. (1976). *Phys. Rev. B*, **13**, 4274.

Gustavsson, S., Leturcq, R., Ihn, T., Ensslin, K., Driscoll, D.C., and Gossard, A.C. (2007). *Physica E*, **40**, 103.

Gustavsson, S., Leturcq, R., Simovic, B., Schleser, R., Ihn, T., Studerus, P., Ensslin, K., Driscoll, D.C., and Gossard, A.C. (2006). *Phys. Rev. Lett.*, **96**, 076605.

Gustavsson, S., Leturcq, R., Simovic, B., Schleser, R., Ihn, T., Studerus, P., Ensslin, K., Driscoll, D.C., and Gossard, A. C. (2008). *Adv. Solid State Phys.*, **46**, 31.

Habib, B., Tutuc, E., and Shayegan, M. (2007). *Appl. Phys. Lett.*, **90**, 152104.

Hall, E.H. (1879). *Am. J. Math.*, **2**, 287.

Hall, E.H. (1880). *The American Journal of Science*, **19**, 200.

Halperin, B.I. (1982). *Phys. Rev. B*, **25**, 2185.

Hanson, R., Kouwenhoven, L.P., Petta, J.R., Tarucha, S., and Vandersypen, L.M.K. (2007). *Rev. Mod. Phys.*, **79**, 1217.

Hartley, R.V.L. (1928). *The Bell System Technical Journal* (July), 535.

Hashimoto, K., Sohrmann, C., Wiebe, J., Inaoka, T., Meier, F., Hirayama, Y., Romer, R. A., Wiesendanger, R., and Morgenstern, M. (2008). *Phys. Rev. Lett.*, **101**, 256802.

Hatano, T., Stopa, M., Yamaguchi, T., Ota, T., Yamada, K., and Tarucha, S. (2004). *Phys. Rev. Lett.*, **93**, 066806.

Hayashi, T., Fujisawa, T., Cheong, H.D., Jeong, Y.H., and Hirayama, Y. (2003). *Phys. Rev. Lett.*, **91**, 226804.

Hedin, L. and Lundqvist, B.I. (1971). *J. Phys. C*, **4**, 2064.

Hedin, L. and Lundqvist, S. (1969). *Solid State Phys.*, **23**, 1.

Heida, J.P., van Wees, B. J., Kuipers, J. J., Klapwijk, T. M., and Borghs, G. (1998). *Phys. Rev. B*, **57**, 11911.

Heinonen, O. (1998). *Composite Fermions: a unified view of the quantum Hall regime*. World Scientific, Singapore.

Heinzel, T. (2007). *Mesoscopic Electronics in Solid State Nanostructures* (2nd edn). Wiley-VCH, Weinheim.

Henny, M., Oberholzer, S., Strunk, C., and Schonenberger, C. (1999). *Phys. Rev. B*, **59**, 2871.

Hikami, S., Larkin, A.I., and Nagaoka, Y. (1980). *Prog. Theor. Phys.*, **63**, 707.

Hofmann, F., Heinzel, T., Wharam, D.A., Kotthaus, J.P., Bohm, G., Klein, W., Trankle, G., and Weimann, G. (1995). *Phys. Rev. B*, **51**, 13872.

Hohenberg, P. and Kohn, W. (1964). *Phys. Rev.*, **136**, B864.

Holleitner, A.W., Decker, C.E., Qin, H., Eberl, K., and Blick, R.H. (2001). *Phys. Rev. Lett.*, **87**, 256802.

Hu, C.-M., Nitta, J., Akazaki, T., Takayanagi, H., Osaka, J., Pfeffer, P., and Zawadzki, W. (1999). *Phys. Rev. B*, **60**, 7736.

Ibach, H. and Luth, H. (1988). *Festkoerperphysik* (2nd edn). Springer, New York.

Ihn, T., Fuhrer, A., Sigrist, M., Ensslin, K., Wegscheider, W., and Bichler, M. (2003). *Advances in Solid State Physics*, **43**, 139.

Imamura, H., Aoki, H., and Maksym, P.A. (1998). *Phys. Rev. B*, **57**, R4257.

Imry, Y. (2002). *Introduction to mesoscopic physics* (2nd edn). Oxford University Press, Oxford.

Ishibashi, K., Takagaki, Y., Gamo, K., Namba, S., Ishida, S., Murase, K., Aoyagi, Y., and Kawabe, M. (1987). *Solid State Commun.*, **64**, 573.

Isihara, A. and Smrčka, L. (1986). *J. Phys. C: Solid State Physics*, **19**, 6777.

Jacak, L., Hawrylak, P., and Wojs, A. (1998). *Quantum dots*. Springer, Berlin.

Jackson, J.D. (1983). *Electrodynamics* (2nd edn). de Gruyter, New York.

Jacobs, T.M. and Giordano, N. (1998). *Superlatt. Microstruct.*, **23**, 635.

Jain, J.K. (1989). *Phys. Rev. Lett.*, **63**, 199.

Jain, J.K. (2000). *Physics Today*, 39, April 2000.

Jaynes, E.T. (1957). *Phys. Rev.*, **106**, 620.

Ji, Y., Chung, Y., Sprinzak, D., Heiblum, M., Mahalu, D., and Shtrikman, H. (2003). *Nature*, **422**, 415.

Johnson, A.C., Marcus, C.M., Hanson, M.P., and Gossard, A.C. (2004). *Phys. Rev. Lett.*, **93**, 106803.

Johnson, A.C., Petta, J. R., Marcus, C. M., Hanson, M. P., and Gossard, A. C. (2005). *Phys. Rev. B*, **72**, 165308.

Johnson, M.B. (1927). *Phys. Rev.*, **29**, 367.

Jones, B. A., Varma, C.M., and Wilkins, J.W. (1988). *Phys. Rev. Lett.*, **61**, 125.

Jönsson, C. (1961). *Z. Phys.*, **161**, 454.

Jusserand, B., Richards, D., Allan, G., Priester, C., and Etienne, B. (1995). *Phys. Rev. B*, **51**, 4707.

Kaplit, M. and Zemel, J.N. (1968). *Phys. Rev. Lett.*, **21**, 212.

Kastner, M.A. (1992). *Rev. Mod. Phys.*, **64**, 849.

Katsnelson, M.I., Novoselov, K.S., and Geim, A.K. (2006). *Nature Physics*, **2**, 620.

Kawaji, S. and Wakabayashi, J. (1976). *Surf. Sci.*, **58**, 238.

Khmel'nitskii, D.E. (1984). *Physica*, **126B**, 235.

Kim, G.T., Park, J.G., Park, Y.W., Muller-Schwanneke, C., Wagenhals, M., and Roth, S. (1999). *Rev. Sci. Instrum.*, **70**, 2177.

Kittel, C. (1970). *Quantentheorie der Festkoerper.* Oldenburg Verlag, Muenchen–Wien.

Kittel, C. (2005). *Introduction to Solid State Physics* (8th edn). Wiley, Hoboken, N.J.

Knap, W., Skierbiszewski, C., Zduniak, A., Litwin-Staszewska, E., Bertho, D., Kobbi, F., Robert, J.L., Pikus, G.E., Pikus, F.G., Iordanskii, S.V., Mosser, V., Zekentes, K., and Lyanda-Geller, Yu. B. (1996). *Phys. Rev. B*, **53**, 3912.

Kobayashi, K., Aikawa, H., Katsumoto, S., and Iye, Y. (2002). *Phys. Rev. Lett.*, **88**, 256806.

Kogan, A., Amasha, S., Goldhaber-Gordon, D., Granger, G., Kastner, M.A., and Shtrikman, H. (2004). *Phys. Rev. Lett.*, **93**, 166602.

Kohn, W. and Luttinger, J.M. (1957). *Phys. Rev.*, **108**, 590.

Kohn, W. and Sham, L.J. (1965). *Phys. Rev.*, **140**, A1133.

Kondo, J. (1964). *Prog. Theor. Phys.*, **32**, 37.

Kondo, J. (1969). In *Solid State Physics*, New York, pp. 183. Academic Press.

Konig, M., Tschetschetkin, A., Hankiewicz, E. M., Sinova, J., Hock, V., Daumer, V., Schafer, M., Becker, C. R., Buhmann, H., and Molenkamp, L. W. (2006). *Phys. Rev. Lett.*, **96**, 076804.

Koppens, F.H.L., Buizert, C., Tielrooij, K.J., Vink, I.T., Nowack, K.C., Meunier, T., Kouwenhoven, L.P., and Vandersypen, L.M.K. (2006). *Nature*, **442**, 766.

Kostial, H., Ihn, Th., Kleinert, P., Hey, R., Asche, M., and Koch, F. (1993). *Phys. Rev. B*, **47**, 4485.

Kouwenhoven, L.P., Marcus, C.M., McEuen, P.L., Tarucha, S., Westervelt, R.M., and Wingreen, N.S. (1997). Electron transport in quantum dots. In *Mesoscopic Electron Transport*, Volume 345 of *Nato ASI Series E: Applied Sciences*, pp. 105–214. Kluwer Academic Publishers, Dordrecht.

Kouwenhoven, L. P., Austing, D.G., and Tarucha, S. (2001). *Rep. Prog. Phys.*, **64**, 701.

Kramer, B., Ohtsuki, T., and Kettemann, S. (2005). *Physics Reports*, **417**, 211.

Kukushkin, I.V., Smet, J.H., von Klitzing, K., and Wegscheider, W. (2002). *Nature*, **415**, 409.

Kukushkin, I.V., v. Klitzing, K., and Eberl, K. (1999). *Phys. Rev. Lett.*, **82**, 3665.

Kulik, I.O. and Shekhter, R.I. (1975). *Sov. Phys. JETP*, **41**, 308.

Kumar, A., Saminadayar, L., Glattli, D.C., Jin, Y., and Etienne, B. (1996). *Phys. Rev. Lett.*, **76**, 2778.

Laikhtman, B. and Altshuler, E.L. (1994). *Ann. Phys.*, **232**, 332.

Lambe, J. and Jaklevic, R.C. (1968). *Phys. Rev.*, **165**, 821.

Lambe, J. and Jaklevic, R.C. (1969). *Phys. Rev. Lett.*, **22**, 1371.

Landau, L. and Lifschitz, E.M. (1962). *Lehrbuch der Theoretischen Physik III, Quantenmechanik* (2nd edn). Akademie Verlag, Berlin.

Landauer, R. (1961). *IBM J. Res. Dev.*, **5**, 183.

Landauer, R. (1989). *J Phys.: Condensed Matter*, **1**, 8099.

Landauer, R. (1993). Information is physical. In *Proc. Workshop on Physics and Computation PhysComp'92*, pp. 1. IEEE Comp. Sci.Press.

Laughlin, R.B. (1983). *Phys. Rev. Lett.*, **50**, 1395.

Laux, S.E., Frank, D.J., and Stern, F. (1988). *Surf. Sci.*, **196**, 101.

Lee, P.A. (1986). *Physica*, **140A**, 169.

Lee, P.A. and Stone, A.D. (1985). *Phys. Rev. Lett.*, **55**, 1622.

Lee, P.A., Stone, A.D., and Fukuyama, H. (1987). *Phys. Rev. B*, **35**, 1039.

Leturcq, R., Graf, D., Ihn, T., Ensslin, K., Driscoll, D.D., and Gossard, A.C. (2004). *Europhys. Lett.*, **67**, 439.

Levinshtein, M., Rumyantsev, S., and Shur, M. (1996). *Handbook Series on Semiconductor Parameters*, Volume 2: Ternary and Quaternary III-V Compounds. World Scientific, Singapore.

Li, G. (2003). *arXiv:0803.4016*.

Lier, K. and Gerhardts, R.R. (1994). *Phys. Rev. B*, **50**, 7757.

Lighthill, M.J. (1964). *Fourier Analysis and Generalised Functions*. Cambridge University Press, Cambridge.

Lindemann, S., Ihn, T., Heinzel, T., Zwerger, W., and Ensslin, K. (2002). *Phys. Rev. B*, **66**, 195314.

Liu, R., Odom, B., Kim, J., Yamamoto, Y., and Tarucha, S. (1996). Volume 3 of *Proc. 23rd Int. Conf. on the Physics of Semicond.*, pp. 2399. World Scientific, Singapore.

Livermore, C., Crouch, C.H., Westervelt, R.M., Campman, K.L., and Gossard, A.C. (1996). *Science*, **274**, 1332.

Lo, H.-K., Popescu, S., and Spiller, T. (1998). *Introduction to Quantum Compuation and Information*. World Scientific, Singapore.

Loss, D. and DiVincenzo, D.P. (1998). *Phys. Rev. A*, **57**, 120.

Loss, D. and Goldbart, P.M. (1992). *Phys. Rev. B*, **45**, 13544.

Loss, D., Goldbart, P., and Balatsky, A.V. (1990). *Phys. Rev. Lett.*, **65**, 1655.

Luo, J., Munekata, H., Fang, F.F., and Stiles, P. J. (1990). *Phys. Rev. B*, **41**, 7685.

Luttinger, J.M. (1951). *Phys. Rev.*, **84**, 814.

Luttinger, J.M. and Kohn, W. (1955). *Phys. Rev.*, **97**, 869.

MacDonald, A.H. (1990). *Phys. Rev. Lett.*, **64**, 220.

Machlup, S. (1954). *J. Appl. Phys.*, **25**, 341.

MacLean, K., Amasha, S., Radu, I.P., Zumbuhl, D.M., Kastner, M.A., Hanson, M.P., and Gossard, A.C. (2007). *Phys. Rev. Lett.*, **98**, 036802.

Madelung, O. (1972). *Festkoerpertheorie*. Springer, Berlin.

Mahan, G.D. (2000). *Many-particle physics* (3rd edn). Kluwer Academic/Plenum Publishers, New York.

Maldague, P.F. (1978). *Surf. Sci.*, **73**, 296.

Mandl, B., Stangl, J., Martensson, T., Mikkelsen, A., Eriksson, J., Karlsson, L.S., Bauer, G., Samuelsson, L., and Seifert, W. (2006). *Nano Lett.*, **6**, 1817.

Martin, J., Akerman, N., Ulbricht, G., Lohmann, T., Smet, J. H., von Klitzing, K., and Yacoby, A. (2008). *Nature Physics*, **4**, 144.

Martin, T. (2005). *arXiv:cond-mat/0501208*.

Martin, Th. and Landauer, R. (1992). *Phys. Rev. B*, **45**, 1742.

Maschke, K., Thomas, P., and Gobel, E.O. (1991). *Phys. Rev. Lett.*, **67**, 2646.

Maxwell, J.C. (1873). *A Treatise on Electricity and Magnetism*, Volume 1. Oxford University Press, New York.

Mayer, H. and Roessler, U. (1991). *Phys. Rev. B*, **44**, 9048.

McEuen, P.L., Alphenaar, B.W., and Wheeler, R.G. (1990). *Surf. Sci.*, **229**, 312.

Meetz, K. and Engl, W.L. (1980). *Elektromagnetische Felder*. Springer Verlag.

Meier, L., Salis, G., Shorubalko, I., Gini, E., Schon, S., and Ensslin, K. (2007). *Nature Physics*, **3**, 650.

Meijer, F.E., Morpurgo, A.F., and Klapwijk, T.M. (2002). *Phys. Rev. B*, **66**, 033107.

Meijer, F.E., Morpurgo, A.F., Klapwijk, T.M., Koga, T., and Nitta, J. (2004). *Phys. Rev. B*, **69**, 035308.

Melinte, S., Freytag, N., Horvatic, M., Berthier, C., Levy, L.P., Bayot, V., and Shayegan, M. (1999). *Phys. Rev. Lett.*, **84**, 354.

Merkt, U., Huser, J., and Wagner, M. (1991). *Phys. Rev. B*, **43**, 7320.

Mermin, N.D. (2007). *Quantum Computer Science. An Introduction.* Cambridge University Press, Cambridge.

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

Miller, W.H. (1968). *J. Chem. Phys.*, **48**, 1651.

Millikan, R.A. (1911). *Phys. Rev.*, **32**, 349.

Molenkamp, L.W., Staring, A.A.M., Beenakker, C.W.J., Eppenga, R., Timmering, C.E., Williamson, J.G., Harmans, C.J.P.M., and Foxon, C.T. (1990). *Phys. Rev. B*, **41**, 1274.

Möllenstedt, G. and Duker, H. (1956). *Z. Phys.*, **145**, 377.

Molnar, B., Peeters, F.M., and Vasilopoulos, P. (2004). *Phys. Rev. B*, **69**, 155335.

Morpurgo, A.F., Heida, J.P., Klapwijk, T.M., van Wees, B.J., and Borghs, G. (1998). *Phys. Rev. Lett.*, **80**, 1050.

Naaman, O. and Aumentado, J. (2006). *Phys. Rev. Lett.*, **96**, 100201.

Nagaev, K.E. (1992). *Phys. Lett. A*, **169**, 103.

Nayak, C., Simon, S.H., Stern, A., Freedman, M., and Sarma, S. Das (2008). *Rev. Mod. Phys.*, **80**, 1083.

Nazarov, Y.V. (2003). *Noise in Mesoscopic Physics*. Kluwer Academic Publishers, Dordrecht.

Nazmitdinov, R.G., Simonovic, N.S., and Rost, J.M. (2002). *Phys. Rev. B*, **65**, 155307.

Neder, I., Heiblum, M., Mahalu, D., and Umansky, V. (2007*a*). *Phys. Rev. Lett.*, **98**, 036803.

Neder, I., Ofek, N., Chung, Y., Heiblum, M., Mahalu, D., and Umansky, V. (2007*b*). *Nature*, **448**, 333.

Ng, T.K. and Lee, P.A. (1988). *Phys. Rev. Lett.*, **61**, 1768.

Nielsen, M.A. and Chuang, I.L. (2000). *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge.

Nitta, J., Akazaki, T., Takayanagi, H., and Enoki, T. (1997). *Phys. Rev. Lett.*, **78**, 1335.

Nitta, J. and Koga, T. (2003). *J. Supercond.*, **16**, 689.

Nitta, J., Koga, T., and Takayanagi, H. (2002). *Physica E*, **12**, 753.

Nitta, J., Meijer, F., Narita, Y., and Takayanagi, H. (2000). *Physica E*, **6**, 318.

Nitta, J., Takayanagi, H., and Calvet, S. (1999). *Microelectron. Eng.*, **47**, 85.

Nockel, J.U. and Stone, A.D. (1994). *Phys. Rev. B*, **50**, 17415.

Novak, V., Hirschinger, J., Niedernostheide, F.J., Prettl, W., Cukr, M., and Oswald, J. (1998). Direct experimental observation of the Hall angle in the low-temperature breakdown regime of n-gaas. *Phys. Rev. B*, **58**, 13099.

Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Katsnelson, M.I., Grigorieva, I.V., Dubonos, S.V., and Firsov, A.A. (2005). *Nature*, **438**, 197.

Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Zhang, Y., Dubonos, S.V., Grigorieva, I.V., and Firsov, A.A. (2004). *Science*, **306**, 666.

Novoselov, K.S., Jiang, Z., Zhang, Y., Morozov, S.V., Stormer, H.L., Zeitler, U., Maan, J.C., Boebinger, G.S., Kim, P., and Geim, A.K. (2007). *Science*, **315**, 1379.

Nowack, K.C., Koppens, F.H.L., Nazarov, Yu.V., and Vandersypen, L.M.K. (2007). *Science*, **318**, 1430.

Nyquist, H. (1928). *Phys. Rev.*, **32**, 110.

Oberholzer, S., Sukhorukov, E.V., and Schonenberger, C. (2002). *Nature*, **415**, 765.

Ono, K., Austing, D.G., Tokura, Y., and Tarucha, S. (2002). *Science*, **297**, 1313.

Pan, W., Xia, J.-S., Shvarts, V., Adams, D.E., Stormer, H.L., Tsui, D.C., Pfeiffer, L.N., Baldwin, K.W., and West, K.W. (1999). *Phys. Rev. Lett.*, **83**, 3530.

Pawlak, M. (1994). *IEEE Transactions on Information Theory*, **40**, 1490.

Payne, C. (1989). *J Phys.: Condensed Matter*, **1**, 4931.

Perdew, J.P. and Zunger, A. (1981). *Phys. Rev. B*, **23**, 5048.

Petroff, P.M., Lorke, A., and Imamoglu, A. (2001). *Physics Today*, **54**, 46.

Petta, J.R., Johnson, A.C., Taylor, J.M., Laird, E.A., Yacoby, A., Lukin, M.D., Marcus, C.M., Hanson, M.P., and Gossard, A.C. (2005). *Science*, **309**, 2180.

Pfeiffer, L., West, K.W., Stormer, H.L., and Baldwin, K.W. (1989). *Appl. Phys. Lett.*, **55**, 1888.

Potok, R.M., Folk, J.A., Marcus, C.M., and Umansky, V. (2002). *Phys. Rev. Lett.*, **89**, 266602.

Potok, R.M., Folk, J.A., Marcus, C.M., Umansky, V., Hanson, M., and Gossard, A.C. (2003). *Phys. Rev. Lett.*, **91**, 016802.

Prange, R.E. and Girvin, S.M. (1988). *The quantum Hall effect*. Springer, New York.

Preskill, J. (1998). Quantum computation. Published online at http://www.theory.caltech.edu/∼preskill.

Pustilnik, M., Glazman, L.I., Cobden, D.H., and Kouwenhoven, L.P. (2001). *Lecture Notes in Physics*, **579**, 3.

Reed, M.A., Bate, R.T., Bradshaw, K., Duncan, W.M., Frensley, W.R., Lee, J.W., and Shih, H.D. (1986). *J. Vac. Sci. Technol. B*, **4**, 358.

Reimann, S.M. (2002). *Rev. Mod. Phys.*, **74**, 1283.

Reznikov, M., Heiblum, M., Shtrikman, H., and Mahalu, D. (1995). *Phys. Rev. Lett.*, **75**, 3340.

Roesler, M., Zimmermann, R., and Richert, W. (1984). *Phys. Stat. Solidi B*, **121**, 609.

Rothstein, J. (1951). *Science*, **114**, 171.

Saminadayar, L., Glattli, D.C., Jin, Y., and Etienne, B. (1997). *Phys. Rev. Lett.*, **79**, 2526.

Sasaki, S., Franceschi, S. De, Elzerman, J.M., van der Wiel, W.G., Eto, M., Tarucha, S., and Kouwenhoven, L.P. (2000). *Nature*, **405**, 764.

Sato, T., Hiruma, K., Shirai, M., Tominaga, K., Haraguchi, K., Katsuyama, T., and Shimada, T. (1995). *Appl. Phys. Lett.*, **77**, 447.

Schedelbeck, G., Wegscheider, W., Bichler, M., and Abstreiter, G. (1997). *Science*, **278**, 1792.

Schiff, L.I. (1949). *Quantum Mechanics* (1st edn). Mc Graw Hill, New York.

Schleser, R., Ihn, T., Ruh, E., Ensslin, K., Tews, M., Pfannkuche, D., Driscoll, D.D., and Gossard, A.C. (2005). *Phys. Rev. Lett.*, **94**, 206805.

Schleser, R., Ruh, E., Ihn, T., Ensslin, K., Driscoll, D.C., and Gossard, A.C. (2004). *Appl. Phys. Lett.*, **85**, 2005.

Schmid, J., Weis, J., Eberl, K., and von Klitzing, K. (2000). *Phys. Rev. Lett.*, **84**, 5824.

Schofield, S.R. (2003). *Phys. Rev. Lett.*, **91**, 136104.

Schottky, W. (1918). *Ann. Phys. (Leipzig)*, **57**, 541.

Schubert, E.F. (1996). *Delta-doping of Semiconductors.* Cambridge University Press, Cambridge.

Schuh, B. (1985). *J. Phys. A: Math. Gen.*, **18**, 803.

Schuster, R., Buks, E., Heiblum, M., Mahalu, D., Umansky, V., and Shtrikman, H. (1997). *Nature*, **385**, 417.

Schwarz, M.P., Wilde, M.A., Groth, S., Grundler, D., Heyn, Ch., and Heitmann, D. (2002). *Phys. Rev. B*, **65**, 245315.

Seeger, K. (2004). *Semiconductor Physics: an Introduction* (9th edn). Springer, Berlin.

Senz, V., Heinzel, T., Ihn, T., Ensslin, K., Dehlinger, G., Grtzmacher, D., and Gennser, U. (2000*a*). *Phys. Rev. B*, **61**, R5082.

Senz, V., Ihn, T., Heinzel, T., Ensslin, K., Dehlinger, G., Grtzmacher, D., and Gennser, U. (2000*b*). *Phys. Rev. Lett.*, **85**, 4357.

Sequoia, S.A., Stillman, G.E., and Wolfe, C.M. (1976). *Thin Solid Films*, **31**, 69.

Shannon, C.E. (1948). *The Bell System Technical Journal*, **27**, 379 and 623.

Shapere, A. and Wilczek, F. (1989). *Geometric Phases in Physics*. World Scientific, Singapore.

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

Shockley, W. (1950). *Electrons and Holes in Semiconductors*. D. van Nostrand Company, Inc., New York.

Shubnikov, L. and de Haas, W.J. (1930*a*). *Leiden Commun.*, **207a**.

Shubnikov, L. and de Haas, W.J. (1930*b*). *Leiden Commun.*, **207c**.

Shubnikov, L. and de Haas, W.J. (1930*c*). *Leiden Commun.*, **207d**.

Shubnikov, L. and de Haas, W.J. (1930*d*). *Leiden Commun.*, **210a**.

Sigrist, M., Ihn, T., Ensslin, K., Loss, D., Reinwald, M., and Wegscheider, W. (2006). *Phys. Rev. Lett.*, **96**, 036804.

Singleton, J. (2001). *Band Theory and Electronic Properties of Solids*. Oxford University Press, New York.

Slater, J.C. (1949). *Phys. Rev.*, **76**, 1592.

Slepian, D. (1976). *Proc. IEEE*, **64**, 292.

Slichter, C.P. (1963). *Principles of magnetic resonance*. Harper & Row, New York.

Smith, T.P., Goldberg, B.B., Stiles, P.J., and Heiblum, M. (1985). *Phys. Rev. B*, **32**, 2696.

Smythe, W.R. (1939). *Static and Dynamic Electricity*. McGraw Hill.

Solomon, P.M., Knoedler, C.M., and Wright, S.L. (1984). *IEEE Elec. Dev. Lett. EDL-*, **5**, 379.

Sprinzak, D., Buks, E., Heiblum, M., and Shtrikman, H. (2000). *Phys. Rev. Lett.*, **84**, 5820.

Steane, A. (1998). *Rep. Prog. Phys.*, **61**, 117.

Steinbach, A.H., Martinis, J.M., and Devoret, M.H. (1996). *Phys. Rev. Lett.*, **76**, 3806.

Stern, A. (1992). *Phys. Rev. Lett.*, **68**, 1022.

Stern, A., Aharonov, Y., and Imry, Y. (1990). *Phys. Rev. A*, **41**, 3436.

Stern, F. (1967). *Phys. Rev. Lett.*, **18**, 546.

Stern, F. and Das Sarma, S. (1984). Electron energy levels in GaAs-$Ga_{1-x}Al_xAs$ heterojunctions. *Phys. Rev. B*, **30**, 840.

Stormer, H.L. (1999). *Rev. Mod. Phys.*, **71**, 875.

Stormer, H., Gossard, A.C., and Wiegmann, W. (1982). *Solid State Commun.*, **41**, 707.

Sturge, M.D. (1962). *Phys. Rev.*, **127**, 768.

Sze, S.M. (1981). *Physics of Semiconductor Devices* (2nd edn). John Wiley and Sons, Inc., New York.

Szilard, L. (1929). *Z. Phys.*, **53**, 840.

Tanaka, Y. and Akera, H. (2006). *Phys. Rev. B*, **53**, 3901.

Tans, S.J., Devoret, M.H., Dai, H.J., Thess, A., Smalley, R.E., Geerlings, L.J., and Dekker, C. (1997). *Nature*, **386**, 474.

Taylor, R.P., Leadbeater, M.L., Wittington, G.P., Main, P.C., Eaves, L., Beaumont, S.P., McIntyre, I., Thoms, S., and Wilkinson, C.D.W. (1988). *Surf. Sci.*, **196**, 52.

Thomas, L.H. (1927). *Phil. Mag.*, **3**, 1.

Thouless, D.J. (1977). *Phys. Rev. Lett.*, **39**, 1167.

Timp, G., Chang, A.M., Cunningham, J.E., Chang, T.Y., Mankievich, P., Behringer, R., and Howard, R.E. (1987). *Phys. Rev. Lett.*, **58**, 2814.

Toffoli, T. (1981). *Mathematical Systems Theory*, **14**, 13.

Tomita, A. and Chiao, R.Y. (1986). *Phys. Rev. Lett.*, **57**, 937.

Tsui, D.C., Stormer, H.L., and Gossard, A.C. (1982). *Phys. Rev. Lett.*, **48**, 1559.

Ulreich, S. and Zwerger, W. (1998). *Europhys. Lett.*, **41**, 117.

Usher, A., Nicholas, R.J., Harris, J.J., and Foxon, C.T. (1990). *Phys. Rev. B*, **41**, 1129.

van der Pauw, L.J. (1958*a*). *Philips Research Reports*, **13**, 1.

van der Pauw, L.J. (1958*b*). *Philips Technical Review*, **20**, 220.

van der Wiel, W.G., Franceschi, S.D., Fujisawa, T., Elzerman, J.M., Tarucha, S., and Kouwenhoven, L.P. (2000). *Science*, **289**, 2105.

van der Wiel, W.G., Franceschi, S. De, Elzerman, J.M., Fujisawa, T., Tarucha, S., and Kouwenhoven, L.P. (2003). *Rev. Mod. Phys.*, **75**, 1.

van Houten, H., Beenakker, C.W.J., Williamson, J.G., Broekaart, M.E.I., and van Loodsrecht, P.H.M. (1989). *Phys. Rev. B*, **39**, 8556.

van Wees, B.J., Kouwenhoven, L.P., Willems, E.M.M., Harmans, C.J.P.M., Mooij, J.E., van Houten, H., Beenakker, C.W.J., Williamson, J.G., and Foxon, C.T. (1991). *Phys. Rev. B*, **43**, 12431.

van Wees, B.J., van Houten, H., Beenakker, C.W.J., Williamson, J.G., Kouwenhoven, L.P., van der Marel, D., and Foxon, C.T. (1988). *Phys. Rev. Lett.*, **60**, 848.

Vedral, V. (2006). *Introduction to Quantum Information Science*. Oxford University Press, New York.

von Klitzing, K. (1996). *Physica Scripta*, **T68**, 21.

von Klitzing, K., Dorda, G., and Pepper, M. (1980). *Phys. Rev. Lett.*, **45**, 494.

Wagner, M., Merkt, U., and Chaplik, A.V. (1992). *Phys. Rev. B*, **45**, 1951.

Wallace, P.R. (1947). *Phys. Rev. B*, **71**, 622.

Walukiewicz, W., Ruda, H.E., Lagowski, J., and Gatos, H.C. (1984). *Phys. Rev. B*, **30**, 4571.

Wannier, G.H. (1937). *Phys. Rev.*, **52**, 191.

Webb, R.A., Washburn, S., Umbach, C.P., and Laibowitz, R.B. (1985). *Phys. Rev. Lett.*, **54**, 2696.

Wegewijs, M.R. and Nazarov, Yu.V. (2001). *arXiv:cond-mat/0103579*.

Wei, H.P., Chang, A.M., Tsui, D.C., and Razeghi, M. (1985). *Phys. Rev. B*, **32**, 7016.

Weinmann, D. and Hausler, W.H. (1994). *Europhys. Lett.*, **26**, 467.

Weinmann, D., Hausler, W.H., and Kramer, B. (1995). *Phys. Rev. Lett.*, **74**, 984.

Weisbuch, C. and Vinter, B. (1991). *Quantum Semiconductor Structures: Fundamentals and Applications*. Academic Press, Boston.

Wharam, D.A., Thornton, T.J., Newbury, R., Pepper, M., Ahmed, H., Frost, J.E.F., Hasko, D.G., Peakock, D.C., Ritchie, D.A., and Jones, G.A.C. (1988). *J. Phys. C*, **21**, L209.

Wildoer, J.W.G., Venema, L.C., Rinzler, A.G., Smalley, R.E., and Dekker, C. (1998). *Nature*, **391**, 59.

Wilhelm, U. and Weis, J. (2000). *Physica E*, **6**, 668.

Willett, R., Eisenstein, J.P., Stormer, H.L., Tsui, D.C., Gossard, A.C., and English, J.H. (1987). *Phys. Rev. Lett.*, **59**, 1776.

Williams, C. and Clearwater, S.H. (1997). *Explorations in Quantum Computing*. Springer, New York.

Williams, R. (1990). *Modern GaAs Processing Methods* (2nd edn). Artech House, Boston.

Williams, R.L., Aers, G.C., Poole, P.J., Lefebvre, J., Chithrani, D., and Lamontagne, B. (2001). *J. Cryst. Growth*, **223**, 321.

Wilson, A.H. (1931*a*). The theory of electronic semiconductors. *Proc. R. Soc. London, Series A*, **133**, 458.

Wilson, A.H. (1931*b*). The theory of electronic semiconductors II. *Proc. R. Soc. London, Series A*, **134**, 277.

Winkler, R. (2003). *Spin–Orbit Coupling Effects in Two-Dimensional Electron and Hole Systems*. Springer, Berlin–New York.

Yacoby, A. and Imry, Y. (1990). *Phys. Rev. B*, **41**, 534.

Yang, M.J., Yang, C.H., and Lyanda-Geller, Y.B. (2004). *Europhys. Lett.*, **66**, 826.

Yau, J.-B., Poortere, E.P. De, and Shayegan, M. (2002). *Phys. Rev. Lett.*, **88**, 146801.

Ye, P.D., Tarucha, S., and Weiss, D. (1999). In *Proc. Int. Conf. Phys. Semicond. (ICPS)*, Singapore. World Scientific.

Yu, E.T., McCaldin, J.O., and McGill, T.C. (1992). Band offsets in semiconductor heterojunctions. In *Solid state physics* (ed. H. Ehrenreich and D. Turnbull), Volume 46, pp. 1–146. Academic Press, San Diego.

Yu, P. (2009). Private communication.

Yu, P.Y. and Cardona, M. (2001). *Fundamentals of Semiconductors* (3rd edn). Springer, Berlin–New York.

Zala, G., Narozhny, B.N., and Aleiner, I.L. (2001). *Phys. Rev. B*, **64**, 214204.

Zhang *et al.*, Y. (2005). *Nature*, **438**, 201.

Zinman, J.M. (1972). *Principles of the Theory of Solids.* Cambridge University Press, London.

Zrenner, A., Beham, E., Stufler, S., Findeis, F., Bichler, M., and Abstreiter, G. (2002). *Nature*, **418**, 612.

Zumbuhl, D.M., Marcus, C.M., Hanson, M.P., and Gossard, A.C. (2004). *Phys. Rev. Lett.*, **93**, 256801.

Zutic, I., Fabian, J., and Sarma, S. Das (2004). *Rev. Mod. Phys.*, **76**, 323.

Index