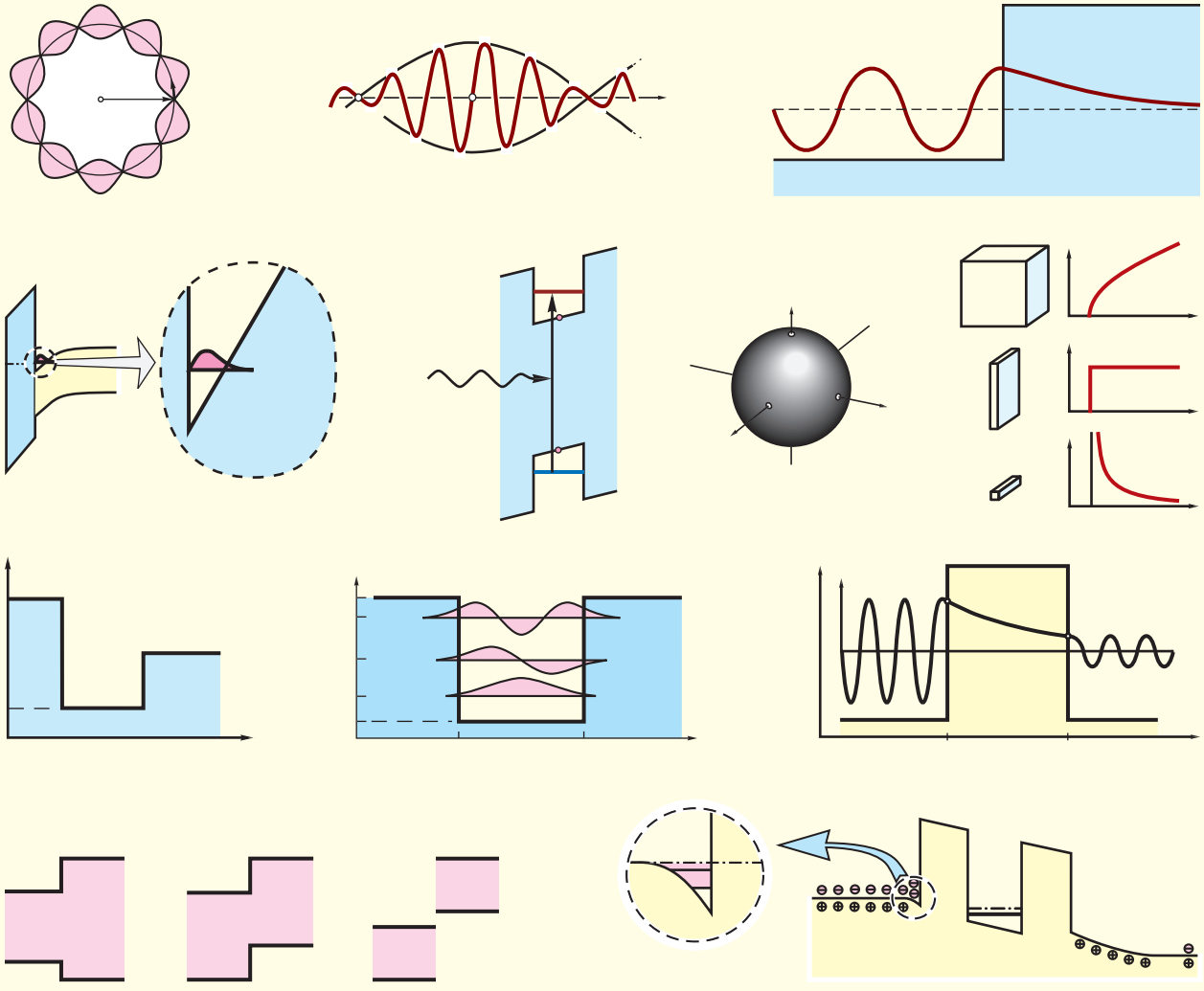
Physical Foundations of Solid-State Devices

E. F. Schubert
Rensselaer Polytechnic Institute
Troy, New York

2006 Edition



This book is recommended as an introductory text for graduate students in Electrical Engineering, Applied Physics, and Material Science.

Physical Foundations of Solid-State Devices

E. F. Schubert
Rensselaer Polytechnic Institute, Troy NY, USA

Introduction

Quantum mechanics plays an essential role in modern semiconductor heterostructure devices. The spatial dimensions of such devices are frequently on the scale of just Angstroms. In the domain of microscopic structures with dimensions comparable to the electron de Broglie wavelength, size quantization occurs. Classical and semi-classical physics no longer gives a correct description of many physical processes. The inclusion of quantum mechanical principles becomes mandatory and provides a most useful description of many physical processes in electronic and photonic heterostructure devices.

Professors, educators, and students in all countries are welcome to use this manuscript as a textbook for their classes. Particularly suited are classes that teach the fundamentals of microelectronic and photonic solid-state devices. I have used this text for several years in a course entitled: "Physical Foundations of Solid-State Devices" that is being taught to beginning graduate and advanced undergraduate students at RPI. The text should be of particular interest to students in Electrical Engineering, Applied Physics, Materials Science, and Microelectronics and Photonics.

I wish to thank all those who have contributed to this text through numerous valuable comments.

E. F. Schubert, December 2006

Table of Contents

1

Classical mechanics and the advent of quantum mechanics

1.1 Newtonian mechanics

The principles of classical mechanics do not provide the correct description of physical processes if very small length or energy scales are involved. ***Classical*** or ***newtonian*** mechanics allows a *continuous* spectrum of energies and allows *continuous* spatial distribution of matter. For example, coffee is distributed homogeneously within a cup. In contrast, quantum mechanical distributions are not continuous but *discrete* with respect to energy, angular momentum, and position. For example, the bound electrons of an atom have discrete energies and the spatial distribution of the electrons has distinct maxima and minima, that is, they are not homogeneously distributed.

Quantum-mechanics does not contradict newtonian mechanics. As will be seen, quantum-mechanics merges with classical mechanics as the energies involved in a physical process increase. In the classical limit, the results obtained with quantum mechanics are identical to the results obtained with classical mechanics. This fact is known as the ***correspondence principle***.

In classical or newtonian mechanics the instantaneous state of a particle with mass m is fully described by the particle's position $[x(t), y(t), z(t)]$ and its *momentum* $[p_x(t), p_y(t), p_z(t)]$. For the sake of simplicity, we consider a particle whose motion is restricted to the x-axis of a cartesian coordinate system. The position and momentum of the particle are then described by $x(t)$ and $p(t) = p_x(t)$. The momentum $p(t)$ is related to the particle's velocity $v(t)$ by $p(t) = m\, v(t) = m\, [\mathrm{d}x(t) / \mathrm{d}t]$. It is desirable to know not only the instantaneous state-variables $x(t)$ and $p(t)$, but also their functional evolution with time. Newton's first and second law enable us to determine this functional dependence. ***Newton's first law*** states that the momentum is a constant, if there are *no forces* acting on the particle, *i. e.*

$$p(t) \;=\; m\, v(t) \;=\; m \frac{\mathrm{d}x(t)}{\mathrm{d}t} \;=\; \text{const.} \tag{1.1}$$

Newton's second law relates an external force, F, to the second derivative of the position $x(t)$ with respect to t,

$$F \;=\; m \frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2} \;=\; m\, a \tag{1.2}$$

where a is the acceleration of the particle. Newton's first and second law provide the state variables $x(t)$ and $p(t)$ in the presence of an external force.

Exercise 1: ***Resistance to Newton's and Kepler's laws***. Newton's laws were greeted with skepticism. Newton postulated that a body continues its uniform motion if there are no forces acting on the body. Opposing contemporaries argued that this would be counter-intuitive and in

utter conflict with daily experience: A carriage must be *constantly pulled* by horses for the carriage to move at a *constant velocity* (see ***Fig***. 1.1). They further argued that as soon as the horses would stop pulling the carriage, it would rest!

Fig. 1.1. Did Newton's first law, which postulates that a body maintains a constant velocity if no forces act on the body, contradict intuition and experience? (after Carriage Association of America, 2004).

What was erroneous in the arguments brought forward by Newton's critics?

Solution: Newton's critics failed to take into account friction forces.

Kepler was greeted with skepticism as well. Not understood by the citizens of the village he was born in, they turned against his mother and accused her of being a witch. Despite the threat of torture, she did not admit to being a witch. Kepler's intervention helped to set her free after she was kept in jail for 14 months. She died six months after her release.

1.2 Energy

Newton's second law is the basis for the introduction of work and energy. Work done by moving a particle along the x axis from 0 to x by means of the force $F(x)$ is defined as

$$W(x) \;=\; \int_0^x F(x)\,\mathrm{d}x \;. \tag{1.3}$$

The energy of the particle *increases* by the (positive) value of the integral given in Eq. (1.3). The total particle energy, E, can be (*i*) purely potential, (*ii*) purely kinetic, or (*iii*) a sum of potential and kinetic energy. If the total energy of the particle is a purely potential energy, $U(x)$, then $W(x) = -U(x)$ and one obtains from Eq. (1.3)

$$F(x) \;=\; -\frac{\mathrm{d}}{\mathrm{d}x}U(x) \;. \tag{1.4}$$

If, on the other hand, the total energy is purely kinetic, Eq. (1.2) can be inserted in the energy equation, Eq. (1.3), and one obtains with $(\mathrm{d}^2/\mathrm{d}t^2)\,x \;=\; v(\mathrm{d}/\mathrm{d}x)\,v$

$$E_{\mathrm{kin}} \;=\; \frac{1}{2}\,m\,v^2 \;=\; \frac{p^2}{2\,m} \;. \tag{1.5}$$

If no external forces act on the particle, then the total energy of the particle is a constant and is the sum of potential and kinetic energy

$$E_{\mathrm{total}} \;=\; E_{\mathrm{kin}} \;+\; U(x) \;=\; \frac{p^2}{2\,m} \;+\; U(x) \;. \tag{1.6}$$

An example of a one-dimensional potential function is shown in ***Fig***. 1.2. Consider a classical object, *e. g*. a soccer ball, positioned on one of the two slopes of the potential, as shown in ***Fig***. 1.2. Once the ball is released, it will move downhill with increasing velocity, reach the maximum velocity at the bottom, and move up on the opposite slope until it comes to a momentary complete stop at the ***classical turning point***. At the turning point, the energy of the ball is purely potential. The ball then reverses its direction of motion and will move again downhill. In the absence of friction, the ball will continue forever to oscillate between the two classical turning points. The total energy of the ball, *i. e*. the sum of potential and kinetic energy remains constant during the oscillatory motion as long as no external forces act on the object.

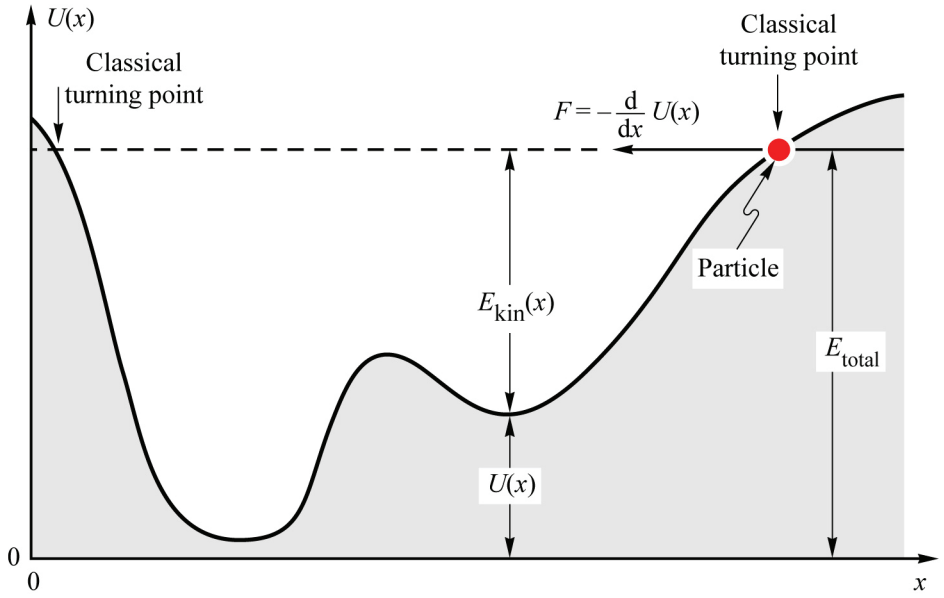


Fig. 1.2. Potential energy as a function of spatial coordinate x. The total energy of the particle shown is the sum of kinetic and potential energies. The force F acting on the particle is the negative derivative of the potential energy with respect to x.

1.3 Hamiltonian formulation of newtonian mechanics

Equations (1.1) and (1.2) are known as the newtonian formulation of classical mechanics. The *hamiltonian formulation* of classical mechanics has the same physical content as the newtonian formulation. However, the hamiltonian formulation focuses on *energy*. The ***hamiltonian function*** $H(x, p)$ is defined as the total energy of a system

$$H(x, p) = \frac{p^2}{2m} + U(x). \tag{1.7}$$

The partial derivatives of the hamiltonian function with respect to x and p are given by

$$\frac{\partial}{\partial x} H(x, p) = \frac{\mathrm{d}}{\mathrm{d}x} U(x) \tag{1.8}$$

U(x) depends on coordinate x
K depends on momentum p

$$\frac{\partial}{\partial p} H(x, p) = \frac{p}{m}. \tag{1.9}$$

Employing these partial derivatives and Eqs. (1.1) and (1.4), one obtains two equations, which are known as the ***hamiltonian equations of motion***:

change in coordinate is velocity

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial}{\partial p} H(x, p) \tag{1.10}$$

momentum changes if there is a force -dU/dx

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial}{\partial x} H(x, p)\,. \tag{1.11}$$

Formally, the linear Eq. (1.1) and the linear, second order differential Eq. (1.2) have been transformed into the *two* linear, first order partial differential Eqs. (1.10) and (1.11). Despite this formal difference, the physical content of the newtonian and the hamiltonian formulation is identical.

1.4 Breakdown of classical mechanics

One of the characteristics of classical mechanics is the ***continuous, non-discrete nature*** of its variables position, $x(t)$, and momentum, $p(t)$. That is, a particle can have any (non-relativistic) momentum with no restrictions imposed by the axioms of classical mechanics. A second characteristic of classical mechanics is the ***deterministic nature*** of time dependent processes. Once initial conditions of a mechanical problem are known (that is the position and momentum of all particles of the system), the subsequent evolution of particle motion can be *predetermined* according to the hamiltonian or newtonian equations of motion. In its final consequence, determinism would predetermine the entire universe from its birth to its death. However, quantum-mechanics shows, that physical processes are not predetermined in a mathematically exact sense. As will be seen later, the determinism inherent to newtonian mechanics is in conflict with with quantum mechanics.

Quantum mechanical principles were first postulated by Planck to explain the black-body radiation. At the end of the 19th century, scientists tried to explain the emission intensity spectrum of a *black body* with temperature T. A black body is defined as a perfectly absorbing, non-reflecting body. The intensity spectrum $I(\lambda)$ (in Watts per unit surface area of the black body and per unit wavelength) was experimentally found to be the same for all black bodies at the same temperature, as predicted by arguments based on thermodynamics. The spectral intensity of black-body radiation as a function of wavelength is shown in ***Fig***. 1.3 for different temperatures.

Rayleigh and Jeans predicted a law based on laws of mechanics, electromagnetic theory and statistical thermodynamics. The Rayleigh-Jeans formula is given by

$$I(\lambda) = \frac{2\,\pi\,k\,T}{\lambda^2} \tag{1.12}$$

where k is Boltzmann's constant. However, this formula yielded agreement with experiments only for long wavelengths. For short wavelengths, namely in the ultraviolet region, the formula has a singularity, *i. e.* $I(\lambda \to 0) \to \infty$. Thus, the formula cannot be physical meaningful and this non-physical phenomenon has been called the "***UV catastrophe***".

Planck (1900) postulated that the oscillating atoms of a black body radiate energy only in the discrete, *i. e.* quantized amounts

$$E = \hbar\omega,\ 2\,\hbar\omega,\ 3\,\hbar\omega,\ \cdots = \hbar c\,\frac{2\pi}{\lambda},\ 2\,\hbar c\,\frac{2\pi}{\lambda},\ 3\,\hbar c\,\frac{2\pi}{\lambda},\ \cdots \tag{1.13}$$

where $\hbar = h/(2\pi)$ is *Planck's constant* divided by 2π, c is the velocity of light, and $\omega = 2\pi f$ is the intrinsic angular frequency of the radiating oscillators. The quantum constant or Planck's

constant has a value of

$$\hbar \;=\; h/(2\pi) \;=\; 1.05 \times 10^{-34}\,\mathrm{Js}\,. \tag{1.12}$$

Employing this postulate in the black-body radiation problem, Planck obtained the following formula for the spectral intensity distribution of a black body at temperature T

$$I(\lambda) \;=\; \frac{4\pi\hbar c^2}{\lambda^5\left[\exp\left(\dfrac{2\pi\hbar c}{\lambda k T}\right) - 1\right]} \tag{1.15}$$

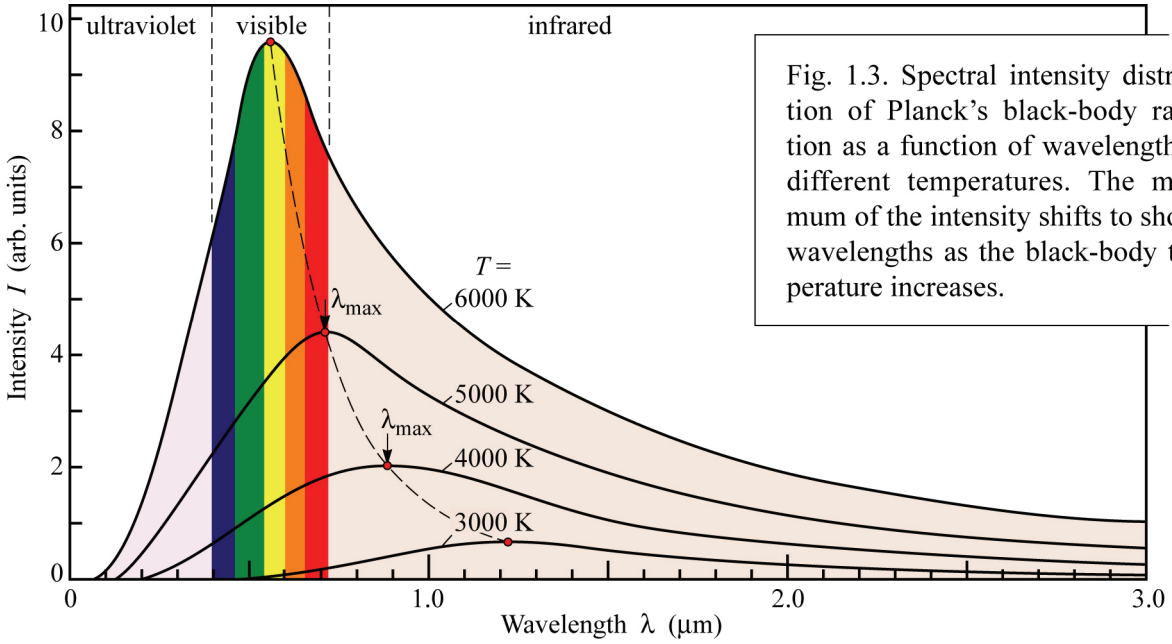


Fig. 1.3. Spectral intensity distribution of Planck's black-body radiation as a function of wavelength for different temperatures. The maximum of the intensity shifts to shorter wavelengths as the black-body temperature increases.

Planck's law of black body radiation was in agreement with experimental observations. The maximum intensity of radiation can be deduced from Eq. (1.15) and it occurs at the wavelengths given by ***Wien's law***

$$\lambda_{\max}\, T \;=\; \text{const.} \;=\; 2880\ \mu\mathrm{m\ K}\,. \tag{1.16}$$

This law predicts that the maximum intensity shifts to the blue spectral region as the temperature of the black body is increased. The energy of the black body radiation at the intensity maximum is given by $E_{\max} = h\,c\,/\,\lambda_{\max}$ and $E_{\max}$ equals about five times the thermal energy, that is $E_{\max} = 4.98\ kT$. Several black body radiation spectra are shown in ***Fig***. 1.3.

Planck's postulate of discrete, *allowed* energies of atomic oscillators as well as of *forbidden*, or *disallowed* energies marks the historical origin of quantum mechanics. It took scientists several decades to come to a complete understanding of quantum mechanics. In the following, the basic postulates of quantum mechanics will be summarized and their implications will be discussed.

Exercise 2: ***The color of hot objects***. If an object gets sufficiently hot, it appears to the human eye, to glow in the red region of the visible spectrum. Assume that the emission spectrum of the hot object peaks at a wavelength of 650 nm. Calculate the temperature of the object.

Solution: The temperature of the hot object is 4431 K.

As the temperature of the object is increased further, the glow changes from the reddish to a yellowish color. At even higher temperatures, the light emitted by the object changes to a white glow. Explain these experimental observations based on the black body radiation.

Solution: As the temperature of the black body increases, the peak wavelength shifts from the infrared into the visible and ultimately into the ultraviolet. At low temperatures, only the short-wavelength tail of the radiation reaches into the visible spectrum and the body therefore appears as red. When the black body is at sufficiently high temperatures, the emission spectrum covers the entire visible spectrum and the body appears as white.

Thermal and night-vision imaging systems can detect the thermal radiation emitted by hot electronic components or by humans, as shown in ***Fig***. 1.4. What is the peak wavelength of planckian emission of a hot electronic component of 100 °C and a human body at 37 °C?

Solution: The peak emission wavelength of the electronic component and human body is 7.7 μm and 9.3 μm, respectively.

Why is a planckian radiator assumed to be perfectly black?

Solution: This assumption is made to ensure that the body does not reflect light. If the body were not perfectly black, it would reflect light thereby appearing in a certain color even if it were at 0 K.

How do we call planckian radiation in ordinary language?

Solution: ***Heat glow*** or ***incandescence***.

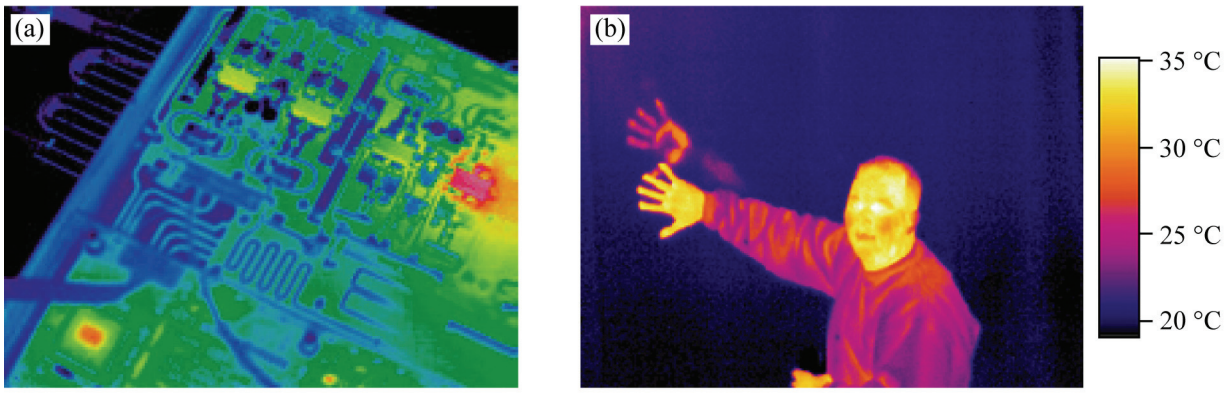


Fig. 1.4. Thermal infrared image obtained in the 3−6 μm wavelength range of (a) an electronic circuit and (b) a person having touched a cold wall and leaving a "thermal imprint" (after Sierra Pacific Corp., 2004; Mikron Corp., 2004).

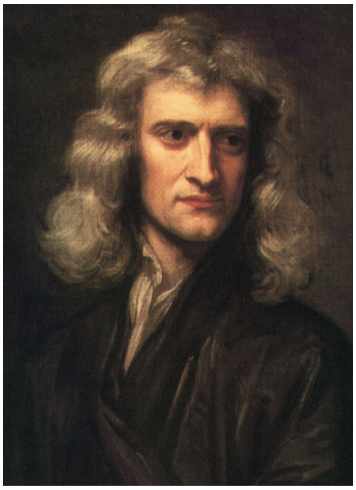
References

Planck M. "Zur Theorie des Gesetzes der Energieverteilung in Normalspectrum" (translated title: "On the theory of the law of energy distribution in normal spectra") *Verhandlungen der Deutschen Gesellschaft* **2**, 237 (1900)

Some general references on quantum mechanics

Bohm D. *Quantum Theory (*Prentice-Hall, Englewood Cliffs, 1951)
Borowitz S. *Fundamental of Quantum Mechanics* (Benjamin, New York, 1967)
Davydov A. S. *Quantum Mechanics*, 2nd edition (Pergamon, Oxford, 1965)
Flügge S. *Practical Quantum Mechanics* Vol. 1 and 2 (Springer Verlag, Berlin, 1971)
Fromhold Jr., A. T. *Quantum Mechanics for Applied Physics and Engineering (*Dover, New York, 1981)
Gillespie D. T., *A Quantum Mechanics Primer (*Int. Textbook Company, Scranton, Pennsylvania, 1970)
Heisenberg W., *The Physical Principles of the Quantum Theory* (Dover, New York, 1949)
Merzbacher E., *Quantum Mechanics*, 2nd edition (John Wiley and Sons, New York, 1970)
Messiah A., *Quantum Mechanics*, Vol. 1 and Vol. 2 (John Wiley and Sons, New York, 1963)
Ridley B. K., *Quantum Processes in Semiconductors (*Clarendon Press, Oxford, 1982)
Saxon D. S. *Elementary Quantum Mechanics (*Holden-Day, San Francisco, 1968)
Sherwin C. W. *Quantum Mechanics* (Holt-Dryden, New York, 1959)
van der Waerden B. L., editor, *Sources of Quantum Mechanics (*Dover, New York, 1967)
Yariv A., *Theory and Applications of Quantum Mechanics (*John Wiley and Sons, New York, 1982)

Johannes Kepler (1571–1630)
Founder of celestial mechanics



Sir Isaac Newton (1642–1727)
Founder of modern mechanics

2

The postulates of quantum mechanics

2.1 The five postulates of quantum mechanics

The formulation of quantum mechanics, also called wave mechanics focuses on the wave function, $\Psi(x, y, z, t)$, which depends on the spatial coordinates x, y, z, and the time t. In the following sections we shall restrict ourselves to one spatial dimension x, so that the wave function depends solely on x. An extension to three spatial dimensions can be done easily. The wave function $\Psi(x, t)$ and its complex conjugate $\Psi^*(x, t)$ are the focal point of quantum mechanics, because they provide a concrete meaning in the macroscopic physical world: The product $\Psi^*(x, t)\ \Psi(x, t)\ \mathrm{d}x$ is the probability to find a particle, for example an electron, within the interval x and $x + \mathrm{d}x$. The particle is described quantum mechanically by the wave function $\Psi(x, t)$. The product $\Psi^*(x, t)\ \Psi(x, t)$ is therefore called the ***window of quantum mechanics to the real world***.

Quantum mechanics further differs from classical mechanics by the employment of ***operators*** rather than the use of ***dynamical variables***. Dynamical variables are used in classical mechanics, and they are variables such as position, momentum, or energy. Dynamical variables are contrasted with ***static variables*** such as the mass of a particle. Static variables do not change during typical physical processes considered here. In quantum mechanics, dynamical variables are replaced by operators which act on the wave function. Mathematical operators are mathematical expressions that act on an operand. For example, $(\mathrm{d}/\mathrm{d}x)$ is the differential operator. In the expression $(\mathrm{d}/\mathrm{d}x)\ \Psi(x, t)$, the differential operator acts on the wave function, $\Psi(x, t)$, which is the operand. Such operands will be used to deduce the quantum mechanical wave equation or Schrödinger equation.

The postulates of quantum mechanics cannot be proven or deduced. The postulates are hypotheses, and, if no violation with nature (experiments) is found, they are called ***axioms***, *i. e.* non-provable, true statements.

Postulate 1

The wave function $\Psi(x, y, z, t)$ describes the temporal and spatial evolution of a quantum-mechanical particle. The wave function $\Psi(x, t)$ describes a particle with one degree of freedom of motion.

Postulate 2

The product $\Psi^*(x, t)\ \Psi(x, t)$ is the probability density function of a quantum-mechanical particle. $\Psi^*(x, t)\ \Psi(x, t)\ \mathrm{d}x$ is the probability to find the particle in the interval between x and $x + \mathrm{d}x$. Therefore,

$$\int_{-\infty}^{\infty} \Psi^*(x, t)\ \ \Psi(x, t)\,\mathrm{d}x \quad = \quad 1 \tag{2.1}$$

If a wave function $\Psi(x, t)$ fulfills Eq. (2.1), then $\Psi(x, t)$ is called a *normalized* wave function. Equation (2.1) is the ***normalization condition*** and implies the fact that the particle must be located somewhere on the x axis.

Postulate 3

The wave function $\Psi(x, t)$ and its derivative $(\partial / \partial x)\, \Psi(x, t)$ are continuous in an isotropic medium.

$$\lim_{x \to x_0} \Psi(x, t) \quad = \quad \Psi(x_0, t) \tag{2.2}$$

$$\lim_{x \to x_0} \frac{\partial}{\partial x} \Psi(x, t) \quad = \quad \frac{\partial}{\partial x} \Psi(x, t) \Big|_{x = x_0} . \tag{2.3}$$

In other words, $\Psi(x, t)$ is a continuous and continuously differentiable function throughout isotropic media. Furthermore, the wave function has to be finite and single valued throughout position space (for the one-dimensional case, this applies to all values of x).

Postulate 4

Operators are substituted for dynamical variables. The operators act an the wave function $\Psi(x, t)$. In classical mechanics, variables such as the position, momentum, or energy are called dynamical variables. In quantum mechanics *operators* rather than dynamical variables are employed. ***Table*** 2.1 shows common dynamical variables and their corresponding quantum-mechanical operators

Dynamical variable in classical mechanics	***Quantum-mechanical operator***	
x	x	(2.4)
$\mathrm{f}(x)$	$\mathrm{f}(x)$	(2.5)
p	$\frac{\hbar}{\mathrm{i}} \frac{\partial}{\partial x}$	(2.6)
$\mathrm{f}(p)$	$\mathrm{f}\left(\frac{\hbar}{\mathrm{i}} \frac{\partial}{\partial x} \right)$	(2.7)
E_{total}	$-\frac{\hbar}{\mathrm{i}} \frac{\partial}{\partial t}$	(2.8)

Table 2.1: Dynamical variables and their corresponding quantum-mechanical operators.

We next substitute quantum mechanical operators for dynamical variables in the total energy equation (see Eq. 1.2.6)

$$\frac{p^2}{2\, m} + U(x) \quad = \quad E_{\mathrm{total}} . \tag{2.9}$$

Using the substitutions of Eqs. (2.4) to (2.8), and inserting the operand $\Psi(x, t)$, one obtains the Schrödinger or quantum mechanical wave equation

$$-\frac{\hbar^2}{2\,m}\frac{\partial^2}{\partial x^2}\Psi(x,t) + U(x)\,\Psi(x,t) = -\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial t}\Psi(x,t)\,. \tag{2.10}$$

The Schrödinger equation is, mathematically speaking, a linear, second order, partial differential equation.

Postulate 5

The expectation value, $\langle\xi\rangle$, of any dynamical variable ξ, is calculated from the wave function according to

$$\boxed{\langle\xi\rangle = \int_{-\infty}^{\infty} \Psi^*(x,t)\,\xi_{\mathrm{op}}\,\Psi(x,t)\,\mathrm{d}x} \tag{2.11}$$

where ξ_{op} is the operator of the dynamical variable ξ. The expectation value of a variable is also referred to as average value or ensemble average, and is denoted by the triangular brackets $\langle\ldots\rangle$. Equation (2.11) allows one to calculate expectation values of important quantities, such as the expectation values for position, momentum, potential energy, kinetic energy, etc.

The five postulates are a concise summary of the principles or quantum mechanics. The postulates have severe implications on the interpretation of ***microscopic*** physical processes. On the other hand, quantum-mechanics smoothly merges into newtonian mechanics for ***macroscopic*** physical processes.

The wave function $\Psi(x, t)$ depends on time. As will be seen in the Section on Schrödinger's equation, the time dependence of the wave function can be separated from the spatial dependence. The wave function can then be written as

$$\Psi(x,t) = \psi(x)\,\mathrm{e}^{\mathrm{i}\,\omega t} \tag{2.12}$$

where $\psi(x)$ is stationary and it depends only on the spatial coordinate. The harmonic time dependence of $\Psi(x, t)$ is expressed by the exponential factor $\exp(\mathrm{i}\,\omega\,t)$.

An example of a stationary wave function is shown in ***Fig***. 2.1 and this wave function is used to illustrate some of the implications of the five postulates. It is assumed that a particle is described by the wave function

$$\psi(x) = A\,(1+\cos x) \qquad \text{for } |x| < \pi \tag{2.13}$$

$$\psi(x) = 0 \qquad \text{for } |x| \geq \pi\,. \tag{2.14}$$

According to the second postulate, the wave function must be normalized, *i. e.*

$$\int_{-\infty}^{\infty} \psi^*(x)\,\psi(x)\,\mathrm{d}x = 1\,. \tag{2.15}$$

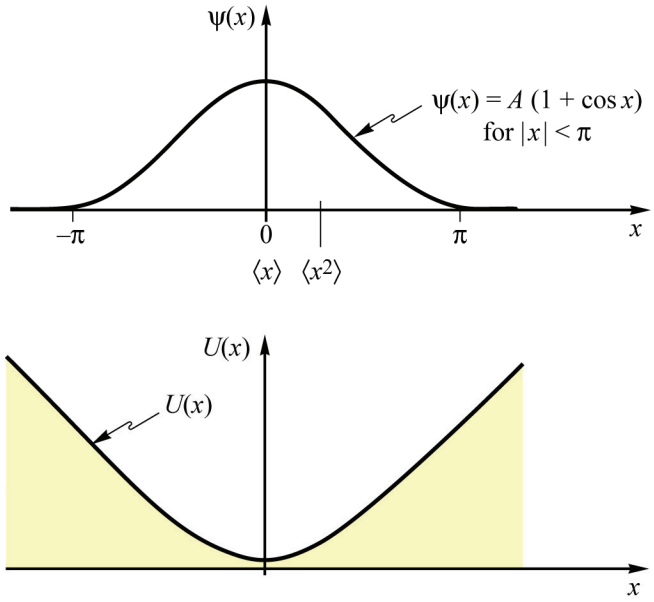


Fig. 2.1. Example for a one-dimensional wave function $\psi(x)$. Also shown is a corresponding potential function, $U(x)$. This potential function provides a driving force towards $x = 0$, that is towards minimum energy.

This condition yields the constant $A = 1/\sqrt{3\pi}$ and thus the normalized wave function is given by

$$\psi(x) = \frac{1}{\sqrt{3\,\pi}}(1 + \cos x) \qquad \text{for } |x| < \pi \tag{2.16}$$

$$\psi(x) = 0 \qquad \text{for } |x| \geq \pi\,. \tag{2.17}$$

Note that $\psi(x)$ is a continuous function and is continuously differentiable throughout position space.

The potential energy of the particle, whose wave function is given by Eqs. (2.16) and (2.17), has a minimum probably around $x = 0$. A guess of such a potential is shown in the lower part of ***Fig***. 2.1. A particle in such a potential experiences a force towards the potential minimum (see Eq. 2.4). Therefore, the corresponding wave function will be localized around the potential-minimum.

Next some expectation values associated with wave function shown in ***Fig***. 2.1 will be calculated using the fifth Postulate. The position expectation value of a particle described by the wave function $\psi(x)$ is given by

$$\langle x \rangle = \int_{-\infty}^{\infty} \psi^*(x)\, x\, \psi(x)\, dx\,. \tag{2.18}$$

Note that x is now an operator, which acts on the wave function $\psi(x)$. Note further that $x\,\psi(x)$ is an odd-symmetry function, and, since $\psi^*(x)$ is an even-symmetry function, the integrand $\psi^*(x)\, x\, \psi(x)$ is again an odd-symmetry function function. The integral over an odd function is zero, *i. e*.

$$\langle x \rangle = 0\,. \tag{2.19}$$

Thus, the expectation value of the position is zero. In other words, the probability to find the particle at any given time is highest at $x = 0$.

It is interesting to know, how far the wave function is distributed from its expectation value. In statistical mathematics, the standard deviation of any quantity, *e. g.* ξ, is defined as

$$\sqrt{\left\langle \xi^2 \right\rangle - \left\langle \xi \right\rangle^2} \quad . \tag{2.20}$$

A measure of the spatial extent of the wave function is the standard deviation of the position of the particle. Hence, the spatial standard deviation of the particle on the *x* axis is given by

$$\sigma \;=\; \sqrt{\left\langle x^2 \right\rangle - \left\langle x \right\rangle^2} \quad . \tag{2.21}$$

With $\langle x \rangle = 0$ one obtains

$$\left\langle x^2 \right\rangle \;=\; \int_{-\pi}^{\pi} \psi^*(x)\, x^2\, \psi(x)\, \mathrm{d}x \;=\; \frac{\pi^2}{3} - \frac{5}{2} \; . \tag{2.22}$$

The standard deviation $\sigma = (\langle x^2 \rangle - \langle x \rangle^2)^{1/2} = (\langle x^2 \rangle)^{1/2}$ is shown in ***Fig***. 2.1 and it is a measure of the spatial extent of the wave function.

The expectation value of the particle momentum can be determined in an analogous way

$$\left\langle p \right\rangle \;=\; \int_{-\infty}^{\infty} \psi^*(x) \left(\frac{\hbar}{\mathrm{i}} \frac{\partial}{\partial x} \right) \psi(x)\, \mathrm{d}x \, . \tag{2.23}$$

Evaluation of the integral yields $\langle p \rangle = 0$. In other words, the particle has no net momentum and it remains spatially at the same location, which is evident for a stationary wave function.

Similarly, the expectation values of kinetic energy, potential energy, and total energy can be calculated if $\psi(x)$ and $U(x)$ are known. The expectation values of these quantities are given by:

Kinetic energy: $$\left\langle E_{\mathrm{kin}} \right\rangle \;=\; \left\langle \frac{p^2}{2m} \right\rangle \;=\; \int_{-\infty}^{\infty} \psi^*(x) \left(\frac{-\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \right) \psi(x)\, \mathrm{d}x \tag{2.24}$$

Potential energy: $$\left\langle U \right\rangle \;=\; \int_{-\infty}^{\infty} \psi^*(x)\, U(x)\, \psi(x)\, \mathrm{d}x \tag{2.25}$$

Total energy:

$$\left\langle E_{\mathrm{total}} \right\rangle \;=\; \left\langle E_{\mathrm{kin}} \right\rangle + \left\langle U \right\rangle \;=\; \int_{-\infty}^{\infty} \psi^*(x) \left[\frac{-\hbar^2}{2\,m} \frac{\partial^2}{\partial x^2} + U(x) \right] \psi(x)\, \mathrm{d}x \tag{2.26}$$

The evaluation of the equations yields that the expectation values of the kinetic, potential, and total energy of the particle are finite and non-zero.

2.2 The de Broglie hypothesis

The de Broglie hypothesis (de Broglie, 1923) is a significant milestone in the development of quantum mechanics because the dualism of waves and matter finds its synthesis in this hypothesis. Typical physical properties that had been associated with matter, before the advent of quantum mechanics, were *mass*, *velocity*, and *momentum*. On the other hand, *wavelength*, *phase-velocity*, and *group-velocity* had been associated with waves. The bridge between the world of waves and the corpuscular world is the de Broglie relation

$$\lambda = h/p \tag{2.27}$$

which attributes a vacuum wavelength λ to a particle with momentum p. This relation, which de Broglie postulated in 1923, can also be written as

$$\boxed{p = \hbar k} \tag{2.28}$$

where $k = 2\pi/\lambda$ is the wavenumber. The kinetic energy of a classical particle can then be expressed in terms of its wavenumber, that is

$$E_{\text{kin}} = \frac{p^2}{2m} = \frac{\hbar^2 k^2}{2m}. \tag{2.29}$$

Four years after de Broglie's hypothesis, Davisson and Germer (1927) demonstrated experimentally that a wavelength can be attributed to an electron, *i. e.* a classical particle. They found, that a beam of electrons with momentum p and wavelength was diffracted by a Ni-crystal the same way as x-rays of the same wavelength λ. The relation between electron momentum p and the x-ray wavelength λ, which yields the *same* diffraction pattern, is given by the de Broglie equation, Eq. (2.27). Thus, a bridge between particles and waves had been built. No longer could one think of electrons as pure particles or x-rays as pure waves. The nature of small particles has both, particle-like and wave-like characteristics. Analogously, a wave has both, wave-like and particle-like characteristics. This fact is known as the dual nature of particles and waves.

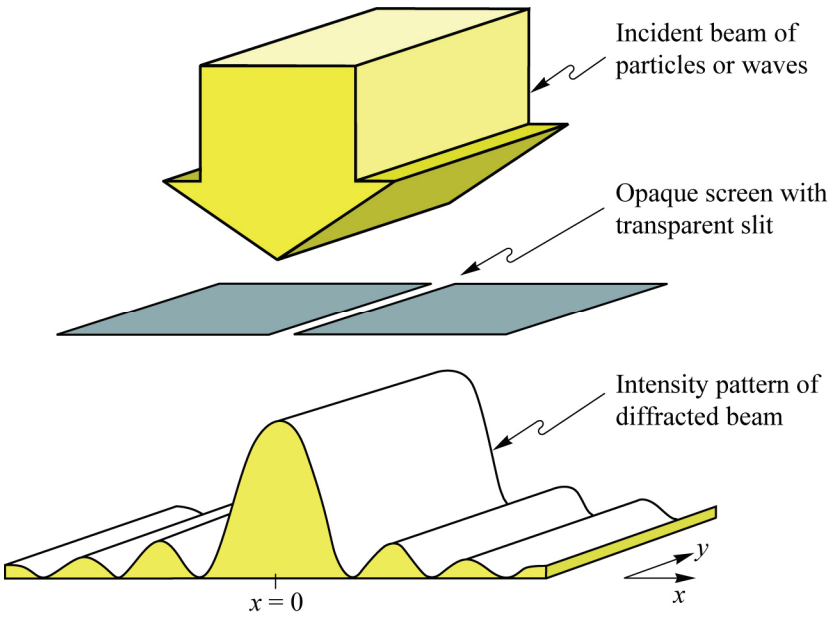


Fig. 2.2. Diffraction of a wave or beam of particles using a transparent slit in a screen. The intensity pattern of the diffracted beam is shown at the bottom.

A simple diffraction experiment is illustrated in ***Fig***. 2.2 which shows a beam of particles incident on a screen having a narrow slit. A diffraction pattern is detected on a screen behind the slit as shown in the lower part. Electrons and x-rays with the same energy generate the same diffraction pattern. The diffraction pattern can be calculated by taking into account the constructive and destructive interference of waves.

The Davisson and Germer experiment further shows that the deterministic nature of classical mechanics is not valid for quantum mechanical particles. No longer is it possible to predict or calculate the exact trajectory of a particle. Instead, one can only calculate *probabilities* (expectation values). For example, the position expectation value of an electron passing through the slit of ***Fig***. 2.2 is $\langle x \rangle = 0$.

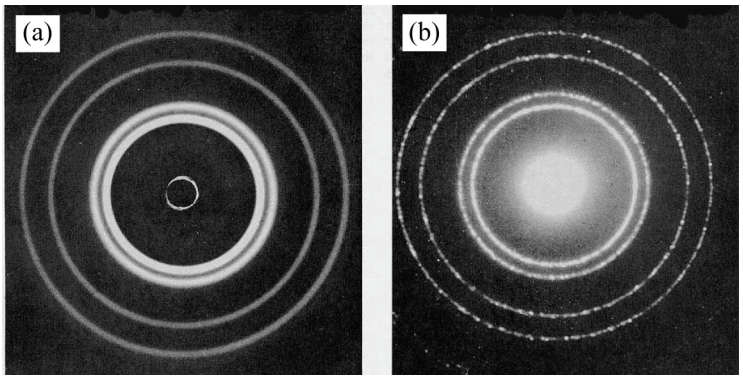


Fig. 2.3. Diffraction pattern of (a) an x-ray beam and (b) an electron beam passing through an Al foil.

Convincing experimental evidence of the wave nature of electrons is shown in ***Fig***. 2.3 which shows two very similar diffraction patterns, one obtained by a beam of x-rays and one by a beam of electrons when passing through an Al foil.

2.3 The Bohr–Sommerfeld quantization condition

During the years 1913–1918, Bohr developed a quantum mechanical model for the electronic states in a hydrogen atom (Bohr, 1913, 1918, 1922). This model supposes that the atom consists of a nucleus with positive charge e and one electron with charge $-e$. The motion of the electron is described by Newton's laws of classical mechanics and a quantum condition. Bohr specifically postulated that an atomic system can only exist in a certain series of electronic states corresponding to a series of discrete values for its energy, and that consequently any change in energy of the system, including the emission and absorption of photons, must take place by a complete transition of the electron between two such states. These states are called as the *stationary* electron states of the system. Bohr further postulated that the radiation absorbed or emitted during a transition between two states possesses a angular frequency ω, given by the relation

$$E_m - E_n = \hbar\omega \tag{2.30}$$

where $\hbar = h/(2\pi)$ is the reduced Planck constant and E_m and E_n are the energies of the two states (the mth state and the nth state) under consideration.

The quantum condition of Bohr can be visualized most easily in terms of the electron de Broglie wave orbiting the nucleus. (Historically, the de Broglie wave concept was postulated in 1925, *i. e.* about a decade *after* the development of Bohr's hydrogen atom model. However, the de Broglie wave concept is used here for convenience). ***Fig***. 2.4 shows a circular electron orbit

of radius r. The electrostatic potential of the nucleus has symmetry and the electron is consequently moving with a constant velocity about the nucleus. Electronic orbitals are allowed, only if the circumference is an integer multiple of the electron de Broglie wavelength

$$S = (n+1)\,\lambda \qquad (n = 0, 1, 2 \ldots) \tag{2.31}$$

where n is an integer and S is the circumference of the electron orbit. If this equation is fulfilled, the electron de Broglie wave is *interfering constructively* with itself as shown in ***Fig***. 2.4(b). Such orbits are called *allowed* orbits. If the latter equation is not fulfilled, the electron wave interferes *destructively* with itself as shown in ***Fig***. 2.4(c). Such orbits are called *forbidden* or *disallowed*.

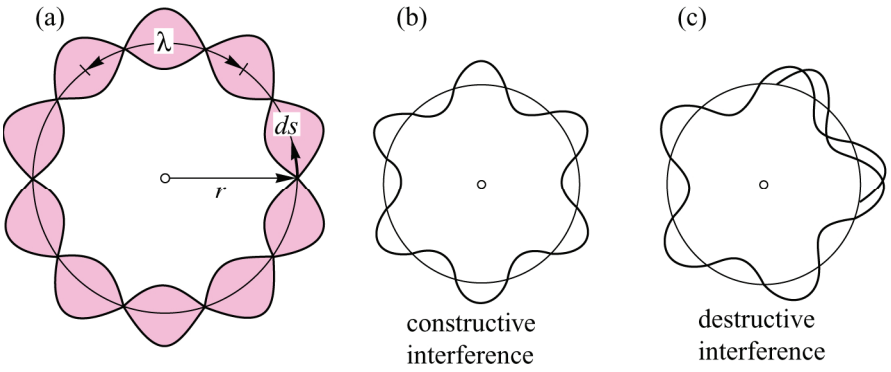


Fig. 2.4. (a) De Broglie wave representing an electron orbiting a nucleus. (b) Constructive inter-ference of wave satisfying the Bohr-Sommerfeld quantum condition. (c) Destructive interference of wave not satisfying the quantum condition results in disallowed state.

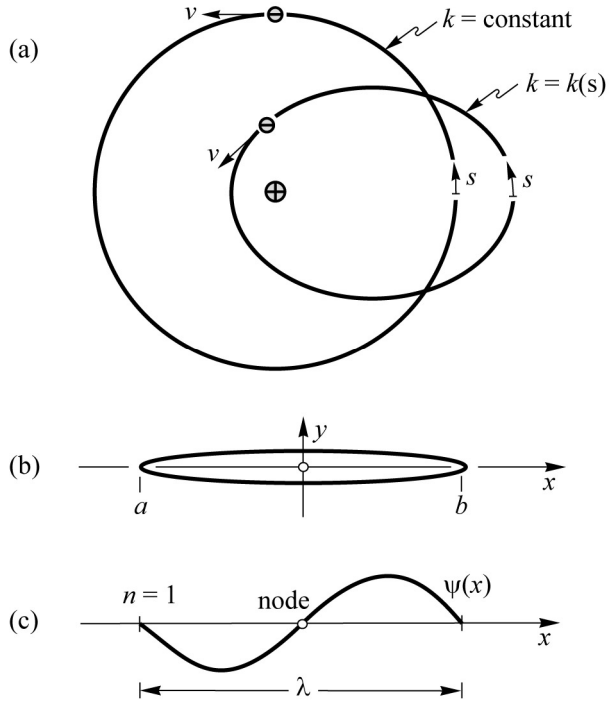


Fig. 2.5. Electrons can orbit the positively charged nucleus on a circular or elliptical curve. (a) The motion of electrons in Bohr's atom model is fully described by (*i*) classical laws (Kepler's laws of planetary motion) and (*ii*) the Bohr-Sommerfeld quantum condition. (b) The one-dimensional Bohr–Sommerfeld quantum condition can be obtained from an ellipsoid compressed onto the x axis. (c) Illustration of quantum state $n = 1$, where the number n is the number of nodes of the wave function.

Only circular orbits have been considered in Eq. (2.31), because the electron is assumed to move in a constant potential with constant momentum $p = h / \lambda$. However, the laws of classical mechanics also allow *elliptical* orbits. For example, the laws of planetary motion (Kepler's laws) allow for elliptical orbits of the planets around the sun. The nucleus is in one of the focal points

of the ellipse as shown in ***Fig***. 2.5. In such elliptical orbits the momentum is a function of the position. It is therefore necessary to generalize the quantum condition of Eq. (2.31) in order to make it applicable to orbits other than circular orbits.

A generalization of the quantum condition is obtained by first rearranging Eq. (2.31) and employing the wavenumber $k = 2\pi / \lambda$ according to

$$\frac{1}{2\pi} k\, S \quad = \quad n \; + \; 1 \qquad\qquad (n = 0,\, 1,\, 2 \ldots) \,. \tag{2.32}$$

Because k depends on the position for elliptical orbits an *integration* rather than a product must be employed

$$\oint_S k(s)\, \mathrm{d}s \quad = \quad 2\pi\,(n + \; 1) \qquad\qquad (n = 0,\, 1,\, 2 \ldots) \,. \tag{2.33}$$

The integral is a closed line integral along the electron orbit S. Using the de Broglie relation $\mathrm{p} = \hbar\, k$ one obtains.

$$\boxed{\oint_S \; p(s)\, \mathrm{d}s \quad = \quad 2\pi\, \hbar\, (n+1) \qquad\qquad (n \; = \; 0, 1, 2 \ldots)} \tag{2.34}$$

which is known as the ***Bohr–Sommerfeld quantization condition***. The integral $\hbar^{-1} \oint p(s) \mathrm{d}s$ is called the phase integral, since it provides the phase change of the electron wave during one complete orbit. The phase integral must have values of multiplies of 2π in order to achieve constructive interference of the electron wave with itself. The properties of the hydrogen atom and of hydrogenic impurities are discussed in greater detail in the Chapter on hydrogenic impurities.

The Bohr–Sommerfeld quantization condition has been derived for a system with three degrees of freedom. In a system with only one degree of freedom, the *one-dimensional Bohr-Sommerfeld condition* applies. To obtain this condition the ellipse shown in ***Fig***. 2.5(b) is compressed to a line on the x axis. Thus, the particle is confined to the x axis. The line-integral of Eq. (2.34) can then be simplified to

$$\oint_S \; p(s)\, \mathrm{d}s \quad = \quad \int_a^b \; p(x)\, \mathrm{d}x \; + \; \int_b^a \; p(x)\, \mathrm{d}x \tag{2.35}$$

Using the fact that the two integrals on the right-hand side of the equation are identical because of symmetry considerations, one obtains the ***one-dimensional Bohr-Sommerfeld quantization condition***

$$\boxed{\int_a^b \; p(x)\, \mathrm{d}x \quad = \quad \pi\, \hbar\, (n+1) \qquad\qquad (n \; = \; 0, 1, 2 \ldots)} \tag{2.36}$$

Most wave functions are oscillating functions. Oscillating functions have locations of zero amplitude, *i. e.* ***nodes***. It is convenient to name the wave functions by the number of nodes. Assume, for example, $n = 1$. Then the phase shift in Eq. (2.36) is 2π. The corresponding wave function has one node. Thus the wave function with the quantum number n has n nodes. The quantum number is identical with the ***number of nodes*** of that wave function.

If we choose $n = 1$ and assume $p(x) = p =$ constant, then, using the de Broglie relation, Eq. (2.36) simplifies to

$$b - a = \lambda . \tag{2.37}$$

The corresponding wave function is shown in ***Fig***. 2.5(c), where the wave function has one node ($n = 1$) in the center. By convention, the nodes at the left and right end of the wave function are not counted.

Exercise 1: ***Bohr's hydrogen atom model***. Many properties of the hydrogen atom can be calculated in terms of the Bohr model. It is based on classical mechanics as well as quantum mechanics. We assume that the electron orbits the hydrogen atom on a circular orbit with radius a_B. The *classical mechanics condition* for the steady state is that the centrifugal force equals the centripetal (coulombic) force, *i. e.*

$$\frac{m\, v^2}{a_B} = \frac{e^2}{4\pi\varepsilon_0\, a_B^2} \tag{2.38}$$

The *quantum mechanical condition* is that the electron wave must interfere constructively with itself (Bohr-Sommerfeld quantization condition), *i. e.*

$$2\pi\, a_B = (n+1)\lambda \qquad \text{for} \quad n = 0, 1, 2 \ldots \tag{2.39}$$

Using Eqs. (2.38) and (2.39) and the de Broglie relation, calculate the Bohr radius, electron potential energy, kinetic energy, and ionization energy (*i. e*. Rydberg energy).

The results of the calculation are:

Bohr radius: $$a_B = (n+1)^2 \frac{4\pi\varepsilon_0\, \hbar^2}{e^2\, m} \tag{2.40}$$

Potential energy: $$E_{pot} = \frac{-1}{(n+1)^2} \frac{e^4\, m}{(4\pi\varepsilon_0\, \hbar)^2} \tag{2.41}$$

Kinetic energy: $$E_{kin} = \frac{1}{2} \frac{1}{(n+1)^2} \frac{e^4\, m}{(4\pi\varepsilon_0\, \hbar)^2} \tag{2.42}$$

Rydberg energy: $$E_{Ryd} = \frac{1}{2} \frac{1}{(n+1)^2} \frac{e^4\, m}{(4\pi\varepsilon_0\, \hbar)^2} \tag{2.43}$$

The hydrogen atom potential, and the potential, kinetic, and Rydberg energy are illustrated in ***Fig***. 2.6. For the ground state of the hydrogen atom, *i. e*. for $n = 0$, one obtains:

$a_B = 0.53$ Å	and	$E_{Ryd} = 13.6$ eV

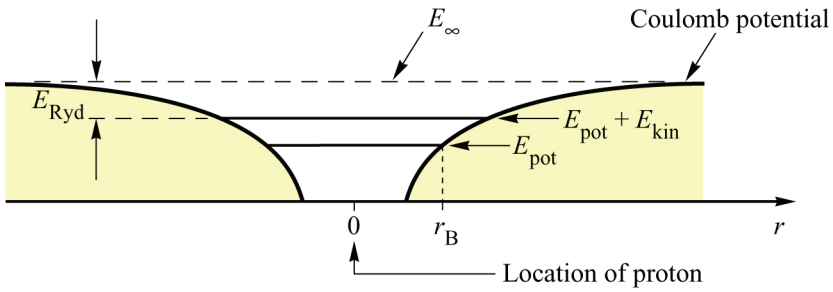


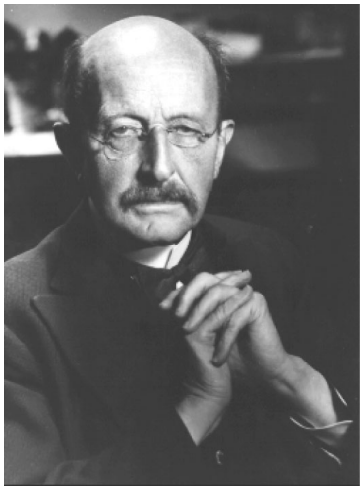
Fig. 2.6. Coulomb potential of a hydrogen atom. The energy of the electron orbiting the proton is the sum of potential and kinetic energy.

Exercise 2: ***Schödinger on quantum rules***. Erwin Schrödinger stated: "The appearance of quantum rules for the hydrogen atom is just as natural as is the existence of resonances for a vibrating string." *Ann. Phys*. **79**, 361 (1926). What are the similarities between the quantum rules of a hydrogen atom and a vibrating string?

Answer: In both cases, only certain modes of oscillation are allowed. For a vibrating string, we call these oscillations that *fundamental oscillation* and its *harmonics*. In the hydrogen atom, we call these oscillations the *ground state* and the *excited states* of the electron.

References

Bohr N. "On the constitution of atoms and molecules" *Philosophical Magazine* **26**, 1 (1913)
Bohr N. "Title unknown to EFS" *Kgl. Danske Vid. Selsk. Skr., nat.-math. Afd., 8. Raekke* **4**, 1 (1918)
Bohr N. "Über die Anwendung der Quantentheorie auf den Atombau I. Die Grundpostulate der Qantentheorie" *Zeitschrift für Physik* 13, 117 (1922)
Davisson C. J. and Germer L. H. "Diffraction of electrons by a crystal of nickel" *Physical Review* **30**, 705 (1927)
de Broglie L. "Waves and quanta" *Nature* **112**, 540 (1923)



Max Planck (1858–1947)
Established equation $E = h\,\nu$

Louis De Broglie (1892–1987)
Attributed wave-like properties to particles

3

Position and momentum space

3.1 Group and phase velocity

Consider a sinusoidal plane wave is propagating along the x axis without any distortion. The wave can be represented by the wave function

$$\Psi(x, t) \quad = \quad \Psi(x - v_{\text{ph}}\, t) \tag{3.1}$$

where v_{ph} is the phase velocity of the wave. This wave possesses translational symmetry, since the wave at the time t is identical to the wave at $t = 0$ shifted on the x axis by an amount of $v_{\text{ph}}\, t$. For example, a sinusoidal plane wave with amplitude A is given by

$$\Psi(x, t) \quad = \quad A \cos (k\, x - \omega\, t). \tag{3.2}$$

The locations of constant phase are given by

$$k\, x \; - \; \omega\, t \quad = \quad \text{const.} \tag{3.3}$$

Differentiation of the position with respect to t yields the ***phase velocity***

$$\boxed{v_{\text{ph}} \quad = \quad \frac{\mathrm{d}x}{\mathrm{d}t} \quad = \quad \frac{\omega}{k}} \tag{3.4}$$

Groups of waves, also called wave packets, can propagate with velocities *different* from the phase velocity. A group of waves is illustrated in ***Fig***. 3.1 by a superposition of two sinusoidal waves with similar angular frequency:

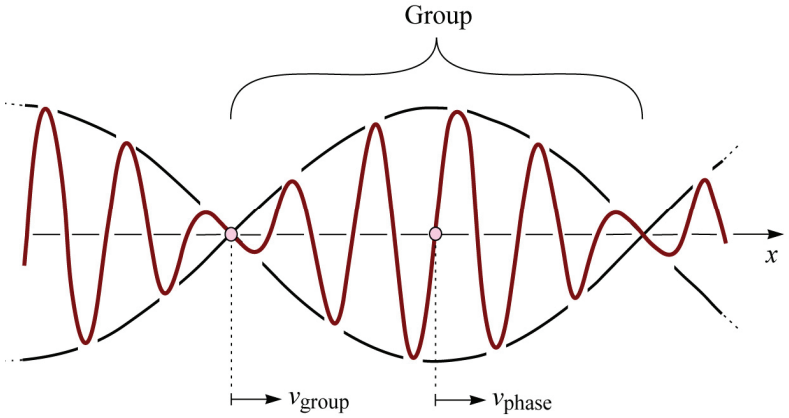


Fig. 3.1. Example for a group of waves propagating along the x-direction. An entire group of wavelets propagates with group velocity v_{group}. Individual wavelets propagate with phase velocity v_{phase}.

$$\Psi(x,t) \;=\; A\cos(k_1\,x-\omega_1\,t) \;+\; A\cos(k_2\,x-\omega_2\,t). \tag{3.5}$$

We define

$$\omega_1 \;=\; \omega-\Delta\omega \qquad \text{and} \qquad \omega_2 \;=\; \omega+\Delta\omega \tag{3.6}$$

$$k_1 \;=\; k-\Delta k \qquad \text{and} \qquad k_2 \;=\; k+\Delta k\ . \tag{3.7}$$

Trigonometric modification of Eq. (3.5) yields

$$\Psi(x,t) \;=\; 2\,A\cos(\omega\,t-k\,x)\;\cos(\Delta\omega\,t-\Delta k\,x)\ . \tag{3.8}$$

With $\Delta\omega << \omega$, we can interpret the wave function as a rapidly oscillating term $\cos(\omega\,t-k\,x)$ and a slowly oscillating term $\cos(\Delta\omega\,t-\Delta k\,x)$ which in turn modulates the amplitude of the rapidly oscillating term. The zeros of the rapidly oscillating term propagate with the phase velocity $v_{ph}=\omega/k$. On the other hand, the phase of the slowly varying term, *i. e.* the wave group, propagates at a velocity $v_{gr}=\Delta\omega/\Delta k$. Thus, for infinitesimal small quantities of $\Delta\omega$ and Δk, we obtain the ***group velocity***

$$\boxed{v_{gr} \;=\; \frac{d\omega}{dk}} \tag{3.9}$$

The group velocity is the velocity at which the wave packets or wavelets propagate in space. The phase velocity can be smaller, equal, or larger than the group velocity. If the phase velocity is larger ($v_{ph} > v_{gr}$), then the wavelets build up at the back end of the group, propagate through the group, and disappear at the front of the group. If the phase velocity is smaller ($v_{ph} < v_{gr}$), then the wavelets are building up at the front end of the group and disappear at the rear end of the group.

In media in which the phase velocity is independent of the frequency of the wave, the phase velocity and group velocity are identical

$$v_{ph} \;=\; v_{gr} \;=\; \frac{\omega}{k} \;=\; \frac{d\omega}{dk}\ . \tag{3.10}$$

Such media are called ***nondispersive media***. Vacuum, and with good approximation also air, are such nondispersive media. The velocity of electromagnetic waves in vacuum and air is $c = 2.99\times10^8$ m/s and this velocity is independent of the frequency of the wave.

If, however, the phase velocity depends on ω, then $v_{ph} \neq v_{gr}$. Media, in which $v_{ph} \neq v_{gr}$ is fulfilled, are called ***dispersive media***. The group velocity in dispersive media can be written as

$$v_{gr} \;=\; v_{ph} \;+\; k\,\frac{dv_{ph}}{dk}\ . \tag{3.11}$$

The validity of this equation can be verified by insertion of $v_{ph}=\omega/k$. Using $k=2\pi/\lambda$ and $d\lambda^{-1} = -\lambda^{-2}\,d\lambda$, one can show that

$$v_{gr} \;=\; v_{ph} \;-\; \lambda\,\frac{dv_{ph}}{d\lambda}\ . \tag{3.12}$$

We have just shown that a well-defined group of waves, i.e. a ***wave packet***, moves with the group velocity. In the classical limit, the group velocity is identical to the propagation velocity of the classical particle described by the wave-packet, *i. e*.

$$v_{\mathrm{gr}} \quad = \quad v_{\mathrm{classical}} \ . \tag{3.13}$$

This requirement is called the ***correspondence principle***. The correspondence principle, which is due to Bohr (1923), postulates a detailed analogy between quantum mechanics and classical mechanics. Specifically it postulates that the results of quantum mechanics merge with those of classical mechanics in the classical limit, *i. e*. for large quantum numbers. Using the definitions of group velocity and of the classical velocity one obtains

$$\frac{\mathrm{d}\omega}{\mathrm{d}k} \quad = \quad \frac{p}{m} \ . \tag{3.14}$$

Substitution of k by using the de Broglie relation ($p = \hbar\, k$) and subsequent integration yields

$$\boxed{E_{\mathrm{kin}} \quad = \quad \hbar\, \omega \quad = \quad \frac{p^2}{2\, m}} \tag{3.15}$$

which is the famous ***Planck relation***. The Planck relation further illustrates the dualism of particles and waves. A particle with momentum p oscillates at an angular frequency ω given by the Planck relation. On the other hand, a wave with angular frequency ω has a momentum p. The kinetic energy $p^2/(2m)$ of the particle coincides with the quantum energy $\hbar\omega$ of the wave representing that particle.

Exercise 1: ***Phase and group velocity***. The experiment described here elucidates the properties of waves, in particular the phase and group velocity. Go to a local pond and throw stones into the water. Watch the water waves created. Several properties of waves can be identified.

The water waves are confined to the surface of the water. What are the curves of constant phase?

Identify the phase and group velocity of the waves. Which of the two velocities is higher? Make a guess for the ratio of $v_{\mathrm{ph}} / v_{\mathrm{gr}}$.

Assume that the distance from the point where a stone enters the water to the shore is x. Can the time it takes for the wave to reach shore be expressed in terms of phase or group velocity and the distance x?

Solution: The curves of constant phase are concentric circles. The phase velocity is higher than the group velocity. Note that individual wavelets appear at the trailing edge of the wave group, move forward through the group of waves, and disappear at the leading edge of the group. The ratio of phase-to-group velocity is given by $v_{\mathrm{ph}} / v_{\mathrm{gr}} \approx 2.0$. The time that it takes the wave to reach the shore is given by $t = x / v_{\mathrm{gr}}$.

3.2 Position-space and momentum-space representation, and Fourier transform

The Fourier transform is a mathematical tool that allows us to represent a function in two different domains. In electrical engineering, we can represent an electrical signal, for example a time-dependent voltage, in the ***time domain*** and the ***frequency domain***. Assume that the voltage

depends on time t and has a quasi-period T (the voltage may not be strictly periodic and thus we use the term "quasi-period"). The Fourier transform of the voltage then depends on the frequency $f = 1/t$ and will have a dominant frequency $f = 1/T$. Finally, the Fourier transform can also be expressed as a function of the *angular frequency* $\omega = 2\pi f$.

Next, we consider the Fourier transform of a wave function $\psi(x)$. Assume that the wave function depends on position x and has a quasi-period λ. Then the Fourier transform of the wave function depends on $1/x$ and has a dominant "frequency" $1/\lambda$. Multiplying $1/\lambda$ by 2π, the Fourier transform can be expressed as a function of the wavenumber $k = 2\pi/\lambda$. Recalling that momentum and wavenumber are related by de Broglie's relation, i.e. $p = \hbar k$, allows us to express the Fourier transform as a function of momentum. Thus a wave function $\psi(x)$ given in ***real space*** has a Fourier transform in ***momentum space***.

According to the 2nd Postulate, the stationary wave function $\psi(x)$ has a clear physical meaning: The probability density is given by the product of the wave function and its complex conjugate, *i. e.* $\psi^*(x)\,\psi(x)$. The wave function $\psi(x)$ can also be represented in momentum space. The momentum-space representation is designated as the ***wave function in momentum space***, $\Phi(p)$. The momentum-space wave function is not a new wave function, but just another representation of a wave function with the same physical content. The two representations are related by the Fourier transform (or Fourier integral). The momentum space representation of the wave function is obtained from the wave function $\psi(x)$ by the ***Fourier transform***

$$\Phi(p) = \frac{1}{\sqrt{2\,\pi\,\hbar}} \int_{-\infty}^{\infty} \psi(x)\, e^{-ipx/\hbar}\, dx \tag{3.16}$$

The wave function in real space is calculated from the wave function in momentum space by the ***inverse Fourier transform***

$$\psi(x) = \frac{1}{\sqrt{2\,\pi\,\hbar}} \int_{-\infty}^{\infty} \Phi(p)\, e^{ipx/\hbar}\, dp \tag{3.17}$$

where $\Phi(p)$ is the amplitude of the momentum space wave function at the momentum p. The Fourier transform provides a unique relationship between the momentum space and position space representation of a particle. That is, for a specific wave function $\psi(x)$ there is only one representation in momentum space $\Phi(p)$.

Another property of the Fourier transform is that the normalization condition holds for the position and momentum space representation. If $\psi(x)$ is normalized, then $\Phi(p)$ is normalized as well.

$$\int_{-\infty}^{\infty} \psi^*(x)\,\psi(x)\,dx = \int_{-\infty}^{\infty} \Phi^*(p)\,\Phi(p)\,dp = 1 \tag{3.18}$$

The Fourier transform (Eqs. 3.16 and 3.17) will not be proven here. The interested reader is referred to the literature (see, for example, Kroemer, 1994). The normalization condition (Eq. 3.18) can be proven by using the Fourier transform.

3.3 Illustrative example: Position and momentum in the infinite square well

A simple quantum-mechanical potential is the square well with infinitely high walls. It will be seen in the Chapter on Schrödinger's equation how to find the wave functions in the infinite

square well potential. Here we are not concerned about how to solve Schrödinger's equation and how to find the wave function. The solutions are shown in ***Fig***. 3.2. The solutions have discrete energies and the wave functions are of sinusoidal shape. The wave function of the lowest state ($n = 0$) and the first excited state ($n = 1$) is given by:

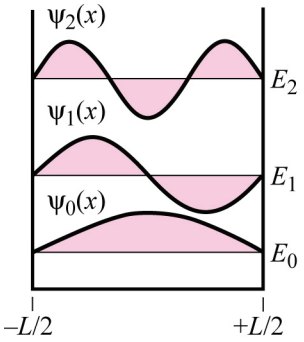
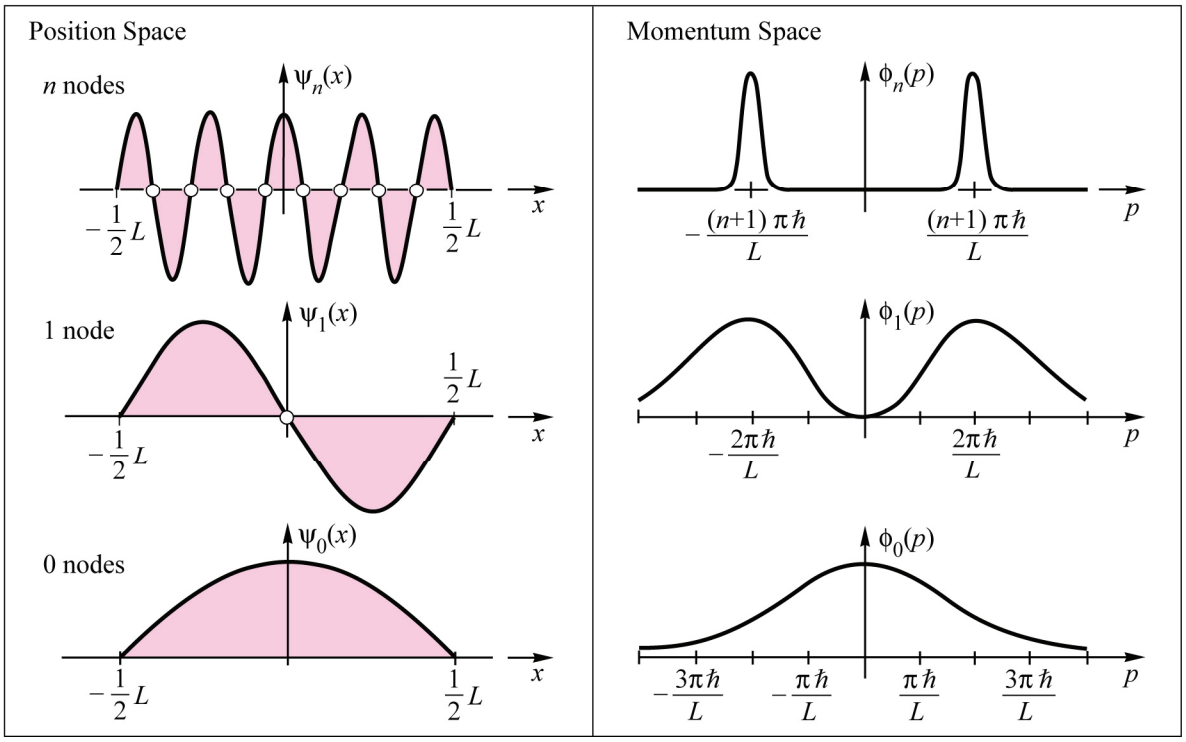


Fig. 3.2. Below: Position space representation and momentum space representation of three wave functions $\psi_0(x)$, $\psi_1(x)$, $\psi_n(x)$. The subscript n refers to the number of nodes (nodes at $x = \pm (1/2)\ L$ are not counted). Left: The three lowest wave functions $\psi_0(x)$, $\psi_1(x)$, $\psi_3(x)$ of an infinite quantum well.



$$\psi_0(x) = \sqrt{\frac{2}{L}}\ \cos\left(\frac{\pi}{L}x\right) \qquad \left(|x| < \frac{1}{2}L\right) \tag{3.19}$$

$$\psi_1(x) = \sqrt{\frac{2}{L}}\ \cos\left(\frac{2\pi}{L}x + \frac{\pi}{2}\right) \qquad \left(|x| < \frac{1}{2}L\right) \tag{3.20}$$

For the nth state, the wave function is given by

$$\psi_n(x) = \sqrt{\frac{2}{L}}\ \cos\left(\frac{(n+1)\pi}{L}x + \frac{n\pi}{2}\right) \quad \left(|x| < \frac{1}{2}L\right) \tag{3.21}$$

Note that these wave functions are normalized. Also note that the wave functions are spatially confined to the region $|x| < (1/2)L$. The magnitude of the wave functions is zero in the barriers, *i. e.*, it is $\psi_n(x) = 0$ for $|x| \geq (1/2)L$.

The Fourier-transform of Eq. (3.16) is used to obtain the momentum space representations of the wave functions. For the wave function with zero nodes, $\psi_0(x)$, one obtains,

$$\begin{aligned} \Phi_0(p) &= \int_{-\infty}^{\infty} \psi_0(x)\, e^{-ipx/\hbar}\, dx \\ &= \sqrt{\frac{L}{\pi\hbar}} \left[\left(\frac{1}{\pi - pL/\hbar} \right) \cos\left(\frac{L}{2\hbar} p \right) + \left(\frac{1}{\pi + pL/\hbar} \right) \cos\left(\frac{L}{2\hbar} p \right) \right]. \end{aligned} \tag{3.22}$$

For a wave function with n nodes, $\psi_n(x)$, one obtains

$$\begin{aligned} \Phi_n(p) &= \int_{-\infty}^{\infty} \psi_n(x)\, e^{-ipx/\hbar}\, dx \\ &= \frac{1/2}{\sqrt{\pi L \hbar}} \left[\frac{1}{i\alpha} \left(i^{n+1}\, e^{-i\frac{pL}{2\hbar}} - i^{-(n+1)}\, e^{i\frac{pL}{2\hbar}} \right) + \frac{1}{i\beta} \left(i^{-(n+1)}\, e^{-i\frac{pL}{2\hbar}} - i^{(n+1)}\, e^{i\frac{pL}{2\hbar}} \right) \right]. \end{aligned} \tag{3.23}$$

where $\alpha = (n+1)(\pi/L) - (p/\hbar)$ and $\beta = -(n+1)(\pi/L) - (p/\hbar)$. The momentum representation $\psi(p)$ is shown for zero, one, and n nodes ($n >> 1$) on the right-hand side of ***Fig***. 3.2. The momentum representation has several interesting aspects. *First*, $\Phi_n(p)$ is a symmetric distribution with respect to p. Consequently, the expectation value of the momentum is zero, since positive and negative momenta compensate one another. The momentum expectation value can be calculated according to the 5th Postulate using the momentum operator given in the 4th Postulate:

$$\langle p \rangle = \int_{-\infty}^{\infty} \psi^*(x)\, \frac{\hbar}{i} \frac{d}{dx}\, \psi(x)\, dx\,. \tag{3.24}$$

The evaluation of this integral gives indeed $\langle p \rangle = 0$. *Second*, the momentum representation has two maxima, one at $p = +(n+1)\pi\hbar/L$ and another one at $p = -(n+1)\pi\hbar/L$. The two-maxima are increasingly pronounced with increasing number of nodes of the wave function. We interpret the *standing wave* $\psi_n(x)$ as a *superposition of two waves*, one propagating in *negative* and another one in *positive x* direction. Returning from the wave-oriented viewpoint to the particle-oriented viewpoint, the wave function $\psi_n(x)$ represents a particle that is propagating back and forth (oscillating) between the boundaries $\pm(1/2)L$ on the x axis. The absolute value of the momentum of the oscillating particle is centered at $p = (n+1)\pi\hbar/L$, as stated above. That is, the particle, represented by the wave function, oscillates between the boundaries $+(1/2)L$ and $-(1/2)L$. Note, however, that the expectation value of the momentum of the oscillating particle is $\langle p \rangle = 0$.

To further visualize the properties of the particle, we note that the particle is represented by a sinusoidal wave function which is confined to $-(1/2)L < x < +(1/2)L$, and has n nodes as shown in the lower left part of ***Fig***. 3.2. The wavelength of the particle is given by $\lambda = 2L/(n+1)$. (Strictly speaking, a wavelength can only be attributed to a wave which is

strictly periodic, and which is not confined to a certain region.) The instantaneous momentum of the particle is given by the de Broglie relationship $p = \hbar k = 2\pi\hbar/\lambda$. Inserting the wavelength into this equation yields the momentum of the particle as

$$p = \frac{(n+1)\,\pi\,\hbar}{L} \,. \tag{3.25}$$

This momentum is in fact the maximum of the momentum distribution obtained from the Fourier transform as shown in ***Fig***. 3.2.

The momentum space representation has, as already mentioned, the same physical content as the position space representation. The expectation values of dynamical variables can be calculated not only from the position space representation (5th Postulate), but also in the momentum space representation. That is, the expectation value of the dynamical variable ξ can be also obtained from the momentum space representation

$$\langle \xi \rangle = \int_{-\infty}^{\infty} \Phi^*(p)\, \xi_{\rm op}\, \Phi(p)\, {\rm d}p \tag{3.26}$$

where $\xi_{\rm op}$ is the operator corresponding to the dynamical variable ξ. We proceed to prove this equation for one specific variable, namely the momentum p. We start with the 3rd and 5th Postulate to determine the momentum expectation value

$$\langle p \rangle = \int_{-\infty}^{\infty} \psi^*(x) \left(\frac{\hbar}{\rm i} \frac{\rm d}{{\rm d}x} \right) \psi(x)\, {\rm d}x \,. \tag{3.27}$$

The wave functions can be represented in momentum space using the Fourier transform of Eq. (3.17):

$$\langle p \rangle = \frac{1}{2\,\pi\,\hbar} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi^*(p')\, {\rm e}^{-{\rm i}p'x/\hbar} \left(\frac{\hbar}{\rm i} \frac{\rm d}{{\rm d}x} \right) \Phi(p)\, {\rm e}^{{\rm i}px/\hbar}\, {\rm d}x\, {\rm d}p\, {\rm d}p' \,. \tag{3.28}$$

The equation can be simplified by differentiating ${\rm e}^{{\rm i}px/\hbar}$ with respect to x, that is

$$\frac{\rm d}{{\rm d}x} {\rm e}^{{\rm i}px/\hbar} = \frac{{\rm i}\,p}{\hbar} {\rm e}^{{\rm i}px/\hbar} \,. \tag{3.29}$$

Introducing this result into Eq. (3.28) yields

$$\langle p \rangle = \frac{1}{2\,\pi\,\hbar} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi^*(p')\, {\rm e}^{-{\rm i}p'x/\hbar}\; p\, \Phi(p)\, {\rm e}^{{\rm i}px/\hbar}\, {\rm d}x\, {\rm d}p\, {\rm d}p' \tag{3.30}$$

The functional dependence of the integrand on x is now known explicitly and therefore the integration over x is done first. Employment of the following relation for the Dirac-delta function,

$$\int_{-\infty}^{\infty} {\rm e}^{{\rm i}(p-p')x/\hbar}\, {\rm d}x = 2\,\pi\,\delta\left(\frac{p-p'}{\hbar} \right) = 2\,\pi\,\hbar\,\delta(p-p') \tag{3.31}$$

yields

$$\langle p \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi^*(p')\, p\, \Phi(p)\, \delta(p - p')\, \mathrm{d}p\, \mathrm{d}p' \ . \tag{3.32}$$

Integration over p' finally yields the momentum expectation value

$$\langle p \rangle = \int_{-\infty}^{\infty} \Phi^*(p)\, p\, \Phi(p)\, \mathrm{d}p \tag{3.33}$$

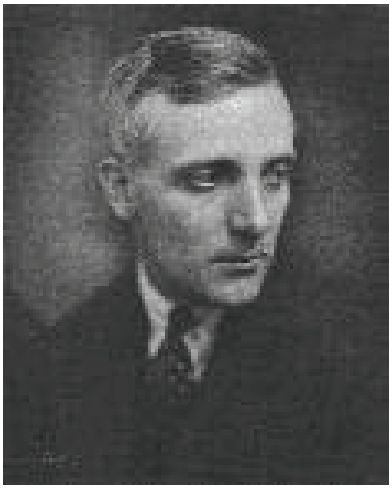
This equation is, for $\xi_{op} = p$, identical with Eq. (3.26), which concludes the proof. The dynamical variable *momentum* corresponds to the operator $(\hbar / \mathrm{i})\,(\mathrm{d} / \mathrm{d}x)$ in position space and to the operator p in momentum space.

What has just been shown for the momentum operator applies to all quantum-mechanical operators: The expectation value of a dynamical variable can be calculated in the position or momentum space representation of the wave function by using the position or momentum space representation of the operator corresponding to the dynamical variable.

References

Bohr N. "The effect of electric and magnetic fields on spectral lines" *Physical Society of London,* **35** 275 (1923)

Kroemer H. *Quantum Mechanics* (Prentice Hall Englewood Cliffs, 1994)



Leon Brillouin (1889–1969)
Author of *Wave Propagation in Periodic Structures* (1946)

4

Operators

4.1 Quantum mechanical operators

Dynamical variables used in classical mechanics are replaced by quantum-mechanical operators in quantum mechanics. Quantum mechanical operators can be used in either position space or momentum space. It was deduced in the preceding section that the dynamical variable *momentum* corresponds to the quantum-mechanical *operator* $(\hbar/\mathrm{i})(\mathrm{d}/\mathrm{d}x)$ in position space, and the variable momentum corresponds to the *operator* p in momentum space. All dynamical variables have quantum-mechanical operators in position and momentum space. Depending on the specific problem it may be more convenient to use either the position-space or momentum-space representation to determine the expectation value of a variable. Table 4.1 summarizes the dynamical variables and their corresponding operators in position and momentum space.

DYNAMICAL VARIABLE		**OPERATOR REPRESENTATION**	
Variable		**Position space**	**Momentum space**
Position	x	x	$-\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial p_\mathrm{x}}$
Potential energy	$U(x)$	$U(x)$	$U\left(-\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial p_\mathrm{x}}\right)$
	$\mathrm{f}(x)$	$\mathrm{f}(x)$	$\mathrm{f}\left(-\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial p_\mathrm{x}}\right)$
Momentum	p_x	$\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x}$	p_x
Kinetic energy	$\frac{p_\mathrm{x}^2}{2m}$	$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}$	$\frac{p_\mathrm{x}^2}{2m}$
	$\mathrm{f}(p_\mathrm{x})$	$\mathrm{f}\left(\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x}\right)$	$\mathrm{f}(p_\mathrm{x})$
Total energy	E_total	$-\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial t}$	$-\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial t}$
Total energy	E_total	$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}+U(x)$	$\frac{p_\mathrm{x}^2}{2m}+U\left(-\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial p_\mathrm{x}}\right)$

Table 4.1: Dynamical variables and corresponding quantum mechanical operators in their position space and momentum space representation. Depending on application it can be more convenient to use either position space or momentum space representation. The function $\mathrm{f}(x)$ denotes any mathematical function of x.

Two operator representations for the total energy are given in ***Table*** 4.1: The first one, $-(\hbar / \mathrm{i})(\partial / \partial t)$, follows from the 4th Postulate. The second one, $p_x^2 / (2\,m) + U(x)$, is the sum of kinetic and potential energy. The specific application determines which of the four operators is most convenient to calculate the total energy.

The total energy operator is an important operator. In analogy to the ***hamiltonian function*** in classical mechanics, the ***hamiltonian operator*** is used in quantum mechanics. The hamiltonian operator thus represents the total energy of the particle represented by the wave function $\psi(x)$

$$H\,\psi(x) \;=\; -\frac{\hbar^2}{2\,m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\psi(x) \;+\; U(x)\,\psi(x) \tag{4.1}$$

or equivalently,

$$H\,\Phi(p) \;=\; \frac{p^2}{2\,m}\,\Phi(p) \;+\; U\!\left(-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}}{\mathrm{d}p}\right)\Phi(p)\;. \tag{4.2}$$

The hamiltonian operator is of great importance because many problems of quantum mechanics are solved by minimizing the total energy of a particle or a system of particles.

4.2 Eigenfunctions and eigenvalues

Any mathematical rule which changes one function into some other function is called an operation. Such an operation requires an ***operator***, which provides the mathematical rule for the operation, and an ***operand*** which is the initial function that will be changed under the operation.

Quantum mechanical operators act on the wave function $\Psi(x, t)$. Thus, the wave function $\Psi(x, t)$ is the operand. Examples for operators are the differential operator $(\mathrm{d} / \mathrm{d}x)$ or the integral operator $\int \ldots \mathrm{d}x$. In the following sections we shall use the symbol ξ_{op} for an operator and the symbol $\mathrm{f}(x)$ for an operand.

The definition of the ***eigenfunction*** and the ***eigenvalue*** of an operator is as follows: If the effect of an operator ξ_{op} operating on a function $\mathrm{f}(x)$ is that the function $\mathrm{f}(x)$ is modified only by the multiplication with a scalar, then the function $\mathrm{f}(x)$ is called the *eigenfunction* of the operator ξ_{op}, that is

$$\xi_{op}\,\mathrm{f}(x) \;=\; \lambda_s\,\mathrm{f}(x) \tag{4.3}$$

where λ_s is a scalar (constant). λ_s is called the ***eigenvalue*** of the eigenfunction. For example, the eigenfunctions of the differential operator are exponential functions, because

$$\frac{\mathrm{d}}{\mathrm{d}x}\,\mathrm{e}^{\lambda_s x} \;=\; \lambda_s\,\mathrm{e}^{\lambda_s x} \tag{4.4}$$

where λ_s is the eigenvalue of the exponential function and the differential operator.

4.3 Linear operators

Virtually all operators in quantum mechanics are *linear* operators. An operator is a linear operator if

$$\xi_{op}\, c\, \psi(x) \quad = \quad c\, \xi_{op}\, \psi(x) \tag{4.5}$$

where c is a constant. For example $\mathrm{d}/\mathrm{d}x$ is a linear operator, since the constant c can be exchanged with the operator $\mathrm{d}/\mathrm{d}x$. On the other hand, the logarithmic operator (log) is not a linear operator, as can be easily verified.

In classical mechanics, dynamical variables obey the ***commutation law***. For example, the product of the two variables *position* and *momentum* commutes, that is

$$x\, p \quad = \quad p\, x \ . \tag{4.6}$$

However, in quantum mechanics the two linear operators, which correspond to x and p, do not commute, as can be easily shown. One obtains

$$x\, p\, \psi(x) \quad = \quad x \left(\frac{\hbar}{\mathrm{i}} \frac{\mathrm{d}}{\mathrm{d}x} \right) \psi(x) \tag{4.7}$$

and alternatively

$$p\, x\, \psi(x) \quad = \quad \frac{\hbar}{\mathrm{i}} \frac{\mathrm{d}}{\mathrm{d}x}\, x\, \psi(x) \quad = \quad \frac{\hbar}{\mathrm{i}}\, \psi(x) \ + \ \frac{\hbar}{\mathrm{i}}\, x \frac{\mathrm{d}}{\mathrm{d}x}\, \psi(x) \ . \tag{4.8}$$

Linear operators do not commute, since the result of Eqs. (4.7) and (4.8) are different.

4.4 Hermitian operators

In addition to linearity, most of the operators in quantum mechanics possess a property which is known as hermiticity. Such operators are hermitian operators, which will be defined in this section. The expectation value of a dynamical variable is given by the 5th Postulate according to

$$\langle \xi \rangle \quad = \quad \int_{-\infty}^{\infty} \psi^*(x)\, \xi_{op}\, \psi(x)\, \mathrm{d}x \ . \tag{4.9}$$

The expectation value $\langle \xi \rangle$ is now assumed to be a physically observable quantity such as position or momentum. Thus, the dynamical variable ξ is real, and ξ is identical to its complex conjugate.

$$\xi \quad = \quad \xi^* \qquad \text{and} \qquad \langle \xi \rangle \quad = \quad \left\langle \xi^* \right\rangle \ . \tag{4.10}$$

It is important to note that $\xi_{op} \neq \xi_{op}^*$. To determine the complex conjugate form of Eq. (4.9) one has to replace each factor of the integrand with its complex conjugate.

$$\left\langle \xi^* \right\rangle \quad = \quad \int_{-\infty}^{\infty} \psi(x)\, \xi_{op}^*\, \psi^*(x)\, \mathrm{d}x \ . \tag{4.11}$$

With Eq. (4.10) one obtains

$$\int_{-\infty}^{\infty} \psi^*(x)\, \xi_{op}\, \psi(x)\, \mathrm{d}x \quad = \quad \int_{-\infty}^{\infty} \psi(x)\, \xi_{op}^*\, \psi^*(x)\, \mathrm{d}x \ . \tag{4.12}$$

Operators which satisfy Eq. (4.12) are called hermitian operators.

The definition of an hermitian operator is in fact more general than given above. In general, ***hermitian operators*** satisfy the condition

$$\boxed{\int_{-\infty}^{\infty} \psi_1^*(x)\, \xi_{\text{op}}\, \psi_2(x)\, \mathrm{d}x \;=\; \int_{-\infty}^{\infty} \psi_2(x)\, \xi_{\text{op}}^*\, \psi_1^*(x)\, \mathrm{d}x} \tag{4.13}$$

where $\psi_1(x)$ and $\psi_2(x)$ may be different functions. If $\psi_1(x)$ and $\psi_2(x)$ are identical, Eq. (4.13) simplifies into Eq. (4.12).

As an example, we consider the observable *variable momentum*. It is easily shown that the momentum operator is an hermitian operator. The momentum expectation value is given by

$$\langle p \rangle \;=\; \int_{-\infty}^{\infty} \psi^*(x)\; \frac{\hbar}{\mathrm{i}} \frac{\mathrm{d}}{\mathrm{d}x}\; \psi(x)\, \mathrm{d}x \;. \tag{4.14}$$

Integration by part (recall: $\int_a^b u'v\,\mathrm{d}x = uv\big|_a^b - \int_a^b uv'\,\mathrm{d}x$) and using $\psi(x \to \infty) = 0$ yields

$$\langle p \rangle \;=\; \int_{-\infty}^{\infty} \psi(x) \left(-\frac{\hbar}{\mathrm{i}} \frac{\mathrm{d}}{\mathrm{d}x} \right) \psi^*(x)\, \mathrm{d}x \tag{4.15}$$

which proves that p is an hermitian operator.

There are a number of consequences and characteristics resulting from the hermiticity of an operator. Two more characteristics of hermitian operators will explicitly mentioned: *First*, ***eigenvalues of hermitian operators are real***. To prove this, suppose ξ_{op} is an hermitian operator with eigenfunction $\psi(x)$ and eigenvalue λ. Then

$$\int_{-\infty}^{\infty} \psi^*(x)\, \xi_{\text{op}}\, \psi(x)\, \mathrm{d}x \;=\; \int_{-\infty}^{\infty} \psi^*(x)\, \lambda\, \psi(x)\, \mathrm{d}x \tag{4.16}$$

$$= \lambda \int_{-\infty}^{\infty} \psi^*(x)\, \psi(x)\, \mathrm{d}x \tag{4.17}$$

and also due to hermiticity of the operator

$$\int_{-\infty}^{\infty} \psi(x)\, \xi_{\text{op}}^*\, \psi^*(x)\, \mathrm{d}x \;=\; \int_{-\infty}^{\infty} \psi(x)\, \lambda^*\, \psi^*(x)\, \mathrm{d}x \tag{4.18}$$

$$= \lambda^* \int_{-\infty}^{\infty} \psi(x)\, \psi^*(x)\, \mathrm{d}x \tag{4.19}$$

Since Eqs. (4.17) and (4.19) are identical, therefore $\lambda = \lambda^*$, which is only true if λ is real. Thus, eigenvalues of hermitian operators are real.

Second, ***eigenfunctions corresponding to two unequal eigenvalues of an hermitian operator are orthogonal to each other***. This is, if ξ_{op} is an hermitian operator and $\psi_1(x)$ and $\psi_2(x)$ are eigenfunctions of this operator and λ_1 and λ_2 are eigenvalues of this operator then

$$\boxed{\int_{-\infty}^{\infty} \psi_1^*(x)\,\psi_2(x)\,\mathrm{d}x \;=\; 0} \quad . \tag{4.20}$$

The two eigenfunctions $\psi_1(x)$ and $\psi_2(x)$ are ***orthogonal*** if they satisfy Eq. (4.20). The statement can be proven by using the hermiticity of the operator ξ_{op}. This yields

$$\int_{-\infty}^{\infty} \psi_1^*(x)\,\xi_{op}\,\psi_2(x)\,\mathrm{d}x \;=\; \int_{-\infty}^{\infty} \psi_2(x)\,\xi_{op}^*\,\psi_1^*(x)\,\mathrm{d}x \;. \tag{4.21}$$

Employing that λ_1 and λ_2 are the eigenvalues of $\psi_1(x)$ and $\psi_2(x)$ yields

$$\lambda_2 \int_{-\infty}^{\infty} \psi_1^*(x)\,\psi_2(x)\,\mathrm{d}x \;=\; \lambda_1 \int_{-\infty}^{\infty} \psi_1^*(x)\,\psi_2(x)\,\mathrm{d}x \;. \tag{4.22}$$

Since λ_1 and λ_2 are unequal, Eq. (4.22) can only be true if $\psi_1(x)$ and $\psi_2(x)$ are orthogonal functions as defined in Eq. (4.20).

4.5 The Dirac bracket notation

A notation which offers the advantage of great convenience was introduced by Dirac (1926). As shown in the proceeding section, wave functions can be represented in position space and, with the identical physical content, in momentum space. Dirac's notation provides a notation which is *independent* of the representation, that is, a notation valid for the position-space and momentum-space representation.

Let $\Psi(x, t)$ be a wave function and let ξ_{op} be an operator; then the following integration is written with the Dirac bracket notation as

$$\left\langle \Psi \middle| \xi_{op} \middle| \Psi \right\rangle \;=\; \int_{-\infty}^{\infty} \Psi^*(x,t)\,\xi_{op}\,\Psi(x,t)\,\mathrm{d}x \tag{4.23}$$

and equivalently in momentum space

$$\left\langle \Psi \middle| \xi_{op} \middle| \Psi \right\rangle \;=\; \int_{-\infty}^{\infty} \Phi^*(p,t)\,\xi_{op}\,\Phi(p,t)\,\mathrm{d}p \;. \tag{4.24}$$

It is important to note the following two points. *First*, because Dirac's notation is valid for the position- and momentum-space representation, the dependences of the wave function on x, t, or p can be left out. Thus, only Ψ and not $\Psi(x, t)$ or $\Psi(p, t)$ may be used in the Dirac notation. However, if desirable, the explicit dependence of Ψ on x, t, or p can be included, for example $\langle\psi(x)\,|\,\xi_{op}\,|\,\psi(x)\rangle$. *Second*, the left-hand-side wave function in the bracket is by definition the complex conjugate of the right-hand-side wave function of the bracket. The integral notation still provides the explicit notation for the complex conjugate wave function, as shown by the asterisk (*).

If the operator equals the *unit-operator* $\xi_{op} = 1$, then

$$\left\langle \Psi \middle| \xi_{op} \middle| \Psi \right\rangle \;=\; \left\langle \Psi \middle| 1 \middle| \Psi \right\rangle \;. \tag{4.25}$$

For convenience the unit operator can be left out

$$\langle \Psi | 1 | \Psi \rangle \;=\; \langle \Psi | 1\,\Psi \rangle \;=\; \langle \Psi | \Psi \rangle \;. \tag{4.26}$$

The normalization condition, given in the 2nd Postulate can then be written as

$$\langle \Psi | \Psi \rangle = \int_{-\infty}^{\infty} \Psi^*(x,t)\,\Psi(x,t)\,\mathrm{d}x \;=\; 1 \;. \tag{4.27}$$

The Dirac notation can also be used to express expectation values. Writing the 5th Postulate in the Dirac's notation, one obtains the expectation value $\langle \xi \rangle$ of a dynamical variable ξ, which corresponds to the operator ξ_{op}, by

$$\langle \xi \rangle \;=\; \langle \Psi | \xi_{op} | \Psi \rangle \;. \tag{4.28}$$

Again, either position-space or momentum space representation of the wave function can be used.

In the Dirac notation, the operator acts on the function on the right hand side of the bracket. To visualize this fact one can write

$$\langle \Psi_1 | \xi_{op} | \Psi_2 \rangle \;=\; \langle \Psi_1 | \xi_{op}\,\Psi_2 \rangle \;. \tag{4.29}$$

If it is required that the operator acts on the first, complex conjugate function, the following notation is used

$$\langle \xi_{op}\,\Psi_1 | \Psi_2 \rangle \;=\; \int_{-\infty}^{\infty} \psi_2(x)\,\xi_{op}^*\,\psi_1^*(x)\,\mathrm{d}x \;. \tag{4.30}$$

Using this notation, the definition for hermiticity of operators reads

$$\langle \Psi_1 | \xi_{op}\,\Psi_2 \rangle \;=\; \langle \xi_{op}\,\Psi_1 | \Psi_2 \rangle \tag{4.31}$$

or equivalently

$$\langle \Psi_1 | \xi_{op} | \Psi_2 \rangle \;=\; \langle \Psi_2 | \xi_{op} | \Psi_1 \rangle^* \;. \tag{4.32}$$

This equation is equivalent to the definition of hermitian operators in Eq. (4.13).

4.6 The Dirac delta function

A valuable function frequently used in quantum mechanics and other fields is the Dirac delta function. The delta function of the variable x is defined as

$$\delta(x - x_0) \;=\; \infty \qquad (x = x_0) \tag{4.33}$$

$$\delta(x - x_0) \;=\; 0 \qquad (x \neq x_0) \;. \tag{4.34}$$

The integral over the delta (δ) function remains finite and the integral has the unit value

$$\int_{-\infty}^{\infty} \delta(x - x_0)\, dx \;=\; 1 \;. \tag{4.35}$$

The δ function can be understood as the limit of a gaussian distribution with an infinitesimally small standard deviation, that is

$$\delta(x - x_0) \;=\; \lim_{\sigma \to 0} \frac{1}{\sigma \sqrt{2\,\pi}} \exp\left[-\frac{1}{2} \left(\frac{x - x_0}{\sigma} \right)^2 \right] . \tag{4.36}$$

Gaussian functions with different standard deviations but the same area under the curve are shown in ***Fig***. 4.1.

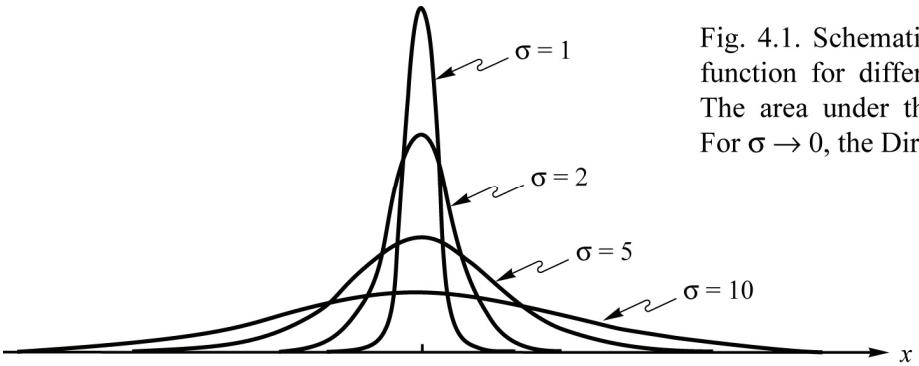


Fig. 4.1. Schematic illustration of a gaussian function for different standard deviations σ. The area under the curve remains constant. For $\sigma \to 0$, the Dirac δ function is obtained.

The δ function can also be represented by its Fourier integral

$$\delta(x - x_0) \;=\; \frac{1}{2\,\pi} \int_{-\infty}^{\infty} e^{i(x - x_0)y}\, dy \;. \tag{4.37}$$

The following equations summarize frequently used properties of the δ function

$$\delta(x) \;=\; \delta(-x) \tag{4.38}$$

$$\delta(\alpha\, x) \;=\; \frac{1}{\alpha}\, \delta(x) \tag{4.39}$$

$$f(x)\, \delta(x - x_0) \;=\; f(x_0)\, \delta(x - x_0) \tag{4.40}$$

$$\int_{-\infty}^{\infty} f(x)\, \delta(x - x_0)\, dx \;=\; f(x_0) \tag{4.41}$$

$$\int_{-\infty}^{\infty} f(x)\, \frac{d}{dx}\, \delta(x - x_0)\, dx \;=\; -\frac{d}{dx}\, f(x) \Big|_{x = x_0} \tag{4.42}$$

References

Dirac P. A. M. "On the theory of quantum mechanics" *Proceedings of the Royal Society* **A109**, 642 (1926)

5

The Heisenberg uncertainty principle

5.1 Definition of uncertainty

Quantum mechanical systems do not allow predictions of their future state with arbitrary accuracy. For example, the outcome of a diffraction experiment such as the Davisson and Germer experiment can be predicted only in terms of a *probability distribution*. It is impossible to predict or calculate the *exact* trajectory of an individual quantum mechanical particle. The Heisenberg uncertainty principle (Heisenberg, 1927) allows us to *quantify the uncertainty associated with quantum mechanical particles*.

The uncertainty of a dynamical variable, $\Delta\xi$, is defined as

$$(\Delta\xi)^2 \;=\; \left\langle \left(\xi \;-\; \langle\xi\rangle\right)^2 \right\rangle . \tag{5.1}$$

Thus, $\Delta\xi$ is the *mean deviation* of a variable ξ from its expectation value $\langle\xi\rangle$. The *mean* deviation can be understood as the most probable deviation. Using the fact that the expectation value of a sum of variables is identical to the sum of the expectation values of these variables, that is

$$\left\langle \sum_i \xi_i \right\rangle \;=\; \sum_i \langle \xi_i \rangle \tag{5.2}$$

one obtains by squaring out Eq. (5.1)

$$(\Delta\xi)^2 \;=\; \left\langle \xi^2 \right\rangle \;-\; \langle\xi\rangle^2 . \tag{5.3}$$

Having *defined* the meaning of uncertainty, we proceed to *quantify* the uncertainty.

5.2 Position–momentum uncertainty

In order to quantify the uncertainty associated with a quantum mechanical system, consider a wave function of gaussian shape as shown in ***Fig***. 5.1. The position space wave function is given by the gaussian function

$$\psi(x) \;=\; \frac{A_x}{\sigma_x \sqrt{2\,\pi}}\; \mathrm{e}^{-\frac{1}{2}\left(\frac{x}{\sigma_x}\right)^2} . \tag{5.4}$$

The constant A_x is used to normalize the wave function using $\langle\psi\,|\,\psi\rangle = 1$, which yields

$$A_x \quad = \quad (4\,\pi)^{1/4}\,\sqrt{\sigma_x}\ . \tag{5.5}$$

This wave function may represent a particle localized in a potential well. If the barriers of the well are sufficiently high the particle cannot escape from the well. That is, the wave function is stationary, *i. e*. $\psi(x)$ does not depend on time.

The momentum distribution which corresponds to the gaussian wave function can be obtained by the Fourier integral

$$\Phi(p) \quad = \quad \frac{1}{\sqrt{2\,\pi\,\hbar}}\ \int_{-\infty}^{\infty}\ \psi(x)\,\mathrm{e}^{-\mathrm{i}px/\hbar}\ \mathrm{d}x. \tag{5.6}$$

Inserting the wave function, given by Eq. (5.4), into the Fourier integral yields

$$\Phi(p) \quad = \quad (4\pi)^{1/4}\ \sqrt{\frac{\hbar}{\sigma_x}}\ \ \frac{1}{\sqrt{2\,\pi}\ \ (\hbar/\sigma_x)}\ \mathrm{e}^{\ -\frac{1}{2}\left(\frac{p}{\hbar/\sigma_x}\right)^2} \qquad . \tag{5.7}$$

This function represents a gaussian function with a prefactor. Thus, the Fourier transform of a gaussian function is also a gaussian function. The gaussian function in Eq. (5.7) has a standard deviation of $\hbar/\sigma_x$ which has the dimension of momentum. Therefore, we introduce the standard deviation in momentum space

$$\sigma_p \quad = \quad \hbar\,/\,\sigma_x\ . \tag{5.8}$$

In analogy to Eq. (5.5), we define the normalization constant A_p as

$$A_p \quad = \quad (4\,\pi)^{1/4}\,\sqrt{\sigma_p}\ . \tag{5.9}$$

Equation (5.7) can be rewritten using the normalization constant A_p.

$$\Phi(p) \quad = \quad \frac{A_p}{\sqrt{2\,\pi}\ \ \sigma_p}\ \mathrm{e}^{\ -\frac{1}{2}\left(\frac{p}{\sigma_p}\right)^2} \qquad . \tag{5.10}$$

This equation is formally identical to the position-space representation of the wave function given by Eq. (5.4). It can be easily verified that the momentum space representation of the state function is normalized as well $\langle\Phi(p)\,|\,\Phi(p)\rangle = 1$. The position and momentum representations of the wave function are shown in ***Fig***. 5.1.

The position space and momentum space representation of a gaussian wave function allows us to form the product of the position and momentum uncertainty using Eq. (5.8)

$$\sigma_x\,\sigma_p \quad = \quad \hbar\ . \tag{5.11}$$

Thus, the product of position uncertainty and momentum uncertainty is a constant. Hence a small position uncertainty results in a large momentum uncertainty and vice versa. The uncertainty of the position Δx, as defined by Eq. (5.1) is, in the case of a gaussian function, identical to the

standard deviation σ_x of that gaussian function. Thus, Eq. (5.11) can be rewritten in its more popular form

$$\Delta x \, \Delta p \quad = \quad \hbar \, . \tag{5.12}$$

This relation was derived for gaussian wave functions and it applies, in a strict sense, only a gaussian wave functions. If the above calculation is performed for wave functions other than a gaussian (*e. g*. square-shaped or sinusoidal), then the uncertainty associated with Δx and Δp is larger. Hence, the general formulation of the ***position–momentum form of the Heisenberg uncertainty relation*** is given by

$$\boxed{\Delta x \, \Delta p \quad \geq \quad \hbar} \tag{5.13}$$

The uncertainty principle shows that an accurate determination of both, position and momentum, cannot be achieved. If a particle is localized on the x axis with a small position uncertainty Δx, then this localization is achieved at the expense of a large momentum uncertainty Δp.

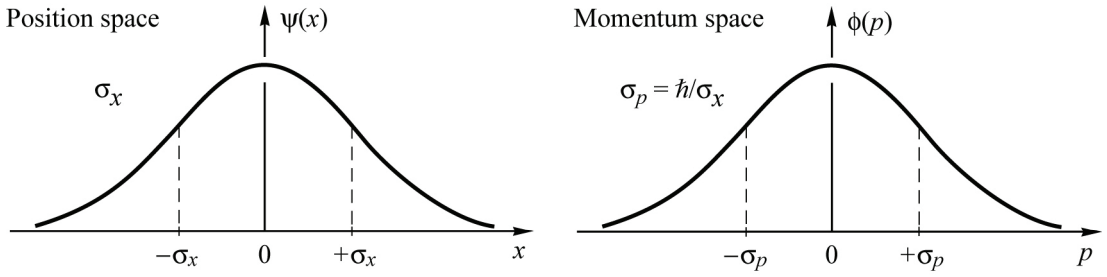


Fig. 5.1. Gaussian wave packet in position space (left). The momentum representation of the Gaussian wavepacket is also a gaussian function (right). The standard deviations are related by $\sigma_p = (h/2\pi) / \sigma_x$. Thus a strongly localized wave packet in position space results in a delocalized function in momentum space and vice versa.

5.3 Energy–time uncertainty

The uncertainty relation between position and momentum will now be modified using the group velocity relation, the de Broglie relation, and the Planck relation to obtain a second uncertainty relation between time and energy. The starting point for this modification is a wave packet that propagates with the group velocity $v_{gr} = \Delta x / \Delta t = \Delta\omega / \Delta k$. Inserting this relation and the de Broglie relation $\Delta p = \hbar \, \Delta k$ into the position-momentum uncertainty relation of Eq. (5.13) yields

$$\Delta t \, \Delta\omega \quad \geq \quad 1 \, . \tag{5.14}$$

It is now straightforward to obtain a second uncertainty relation by employing the Planck relation $\Delta E = \hbar \, \Delta\omega$, which yields

$$\boxed{\Delta E \, \Delta t \quad \geq \quad \hbar} \tag{5.15}$$

which is the ***energy–time form of the Heisenberg uncertainty relation***. This relation states that the energy of a quantum mechanical state can be obtained with highest precision (small ΔE), if

the uncertainty in time is large, *i. e*. for long observation times for quantum-mechanical transitions with a long lifetimes.

The uncertainty relations are valid between the variables x and p (Eq. 5.13) as well as E and t (Eq. 5.15). These pairs of variables are called ***canonically conjugated variables***. A small uncertainty of one variable implies a large uncertainty of the other variable of the *same pair*. On the other hand, the two pairs of variables x, p and E, t are *independent*. For example a large uncertainty in the momentum does not allow any statement about the uncertainty in energy. The uncertainty principle requires that the deterministic nature of classical mechanics be revised. If an uncertainty of momentum or position exists, it is impossible to determine the future trajectory of a particle. On the other hand, the correspondence principle requires that quantum mechanics merges into classical mechanics in the classical limit. Therefore, the uncertainty of Δx or Δp should be insignificantly small in classical physics.

Exercise 1: ***Significance of the uncertainty principle in the macroscopic domain***. To see the insignificance of the uncertainty principle in the macroscopic physical world, a body with mass $m = 1$ kg and velocity $v = 1$ m/s is considered. The body moves along the x axis and the position of its centroid is assumed to be known to an accuracy of $\Delta x_0 = 1$ Å. Calculate the position uncertainty after a time of 10 000 seconds.

Solution: The initial momentum uncertainty is given by $\Delta p_0 \approx \hbar / \Delta x_0 \approx 10^{-24}$ kg m/s. After 10,000 seconds with no forces acting on the body, the position uncertainty is given by

$$\Delta x \;=\; \Delta x_0 + (\Delta p_0 / m)\, t \;\approx\; 1\ \text{Å} + 10^{-10}\ \text{Å}\,. \tag{5.16}$$

This exercise elucidates that the uncertainty principle does not contribute a significant uncertainty in classical mechanics. Thus, even though the trajectory of a macroscopic body cannot be determined in a strict sense, the associated uncertainty is insignificantly small. Even after a time of 10^{18} s (which is 30×10^9 years, *i. e*. approximately the age of the universe) the position uncertainty would be just 1 μm, which is still very small.

Exercise 2: ***The natural linewidth of optical transitions***. Consider a quantum mechanical transition from an excited atomic state to a ground state. Assume that the spontaneous emission lifetime of the transition is τ. The uncertainty relation gives the spectral width of the emission as

$$\Delta E \;=\; \hbar / \tau\,. \tag{5.17}$$

This linewidth is referred to as the *natural linewidth* of a *homogeneously broadened emission line*.

Problem: Assume that the transition probability follows an exponential distribution, as shown in ***Fig***. 5.2. Calculate the spectral shape of the emission line. What is the natural linewidth of a single quantum transition with lifetime 10 ns? Why is the spectral width of a light-emitting diode (LED), typically 50–100 meV at 300 K, much broader than the natural emission linewidth?

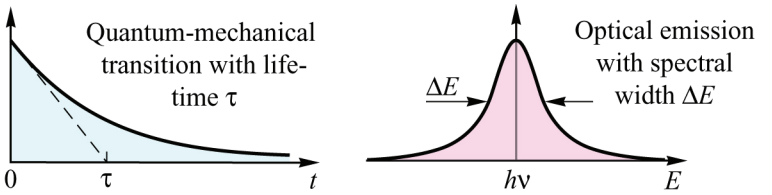


Fig. 5.2. Radiative quantum-mechanical transition with spontaneous lifetime τ and spectral width ΔE.

Solution: The spectral shape of a spontaneous emission line has a lorentzian lineshape given by $I(\nu) = (1/\pi)\,\alpha / [(\nu - \nu_0)^2 + \alpha^2]$. This follows from that fact that the Fourier transform of an exponential function, $\exp(-t/\tau)$, is a lorentzian function. A radiative spontaneous transition with lifetime 10 ns has a spectral width of 6.6×10^{-8} eV. Optical emission in LEDs originates from *many* closely spaced quantum levels distributed over a range of energies in the conduction and valence band. The spectral width of transitions in LEDs is about 2–4 kT (~50–100 meV at 300 K).

Exercise 3: ***The uncertainty relation of information technology***. The uncertainty relation is given by $\Delta t\,\Delta\omega \geq 1$, where "= 1" is valid for gaussian functions and "> 1" is valid for other functions. As an approximation, the following relation, termed the ***uncertainty relation of information technology***, can be used

$$\Delta t\ \Delta f = 1\ . \tag{5.18}$$

The relation applies to the maximum frequency and minimum time duration of electrical pulses. That is, an electronic system that is capable of transmitting a pulse of, say 10 ns, must have a maximum frequency (bandwidth) of Δf = 1/10 ns = 100 MHz. Audio amplifiers have a typical bandwidth of 20 kHz. They can transmit pulses as short as 50 μs.

Problem: Assume that a square-shaped audio pulse of duration 10 μs is being generated. Would a human being be able to hear it (maximum frequency for hearing is 20 kHz)?

Solution: The square pulse would cover a frequency spectrum up to 100 kHz. We would be able to hear the pulse, as frequency components of the square pulse lie within the audio range.

Problem: Very short pulses can be generated by coherent broadband optical source. Assume that a broadband optical signal reaches from λ = 0.75 μm to 1.0 μm. Calculate the Δf corresponding to the signal and the minimum time duration of the optical pulse. How long are the shortest pulses that can be generated?

Solution: Δf = 100 THz and thus Δt = 10 fs. The shortest pulses that currently can be generated by coherent optical broadband sources are < 5 fs.

Exercise 4: ***The uncertainty principle and its meaning to philosophy***. After it became evident that nature is governed by strict laws, philosophers established what was called the *Philosophy of Determinism*. This particular school of philosophy postulated that any event in a system would be determined by the initial conditions and the boundary conditions of the system. Even if we do not know these initial and boundary conditions, the course of events would still be predetermined.

The Philosophy of Determinism has significant implications on human existence. Applying this line of thought to the molecular and the atomic domain would force us to conclude that

human thoughts and decisions also are predetermined. There would be no avenue to influence future events and also the course of our lives. We could stop working, as all future events would be predetermined anyway. Why is the Philosophy of Determinism flawed?

Solution: The discovery of the uncertainty principle by Heisenberg in 1927 removed the basis of the Philosophy of Determinism. The uncertainty is particularly large in the microscopic domain, *e.g.* for molecular reactions in our brain that ultimately govern our decisions and actions. The origins of the Philosophy of Determinism can be traced back to Ancient Greece. The Philosophy of Determinism re-emerged in the AD 1500s, and subsequently gained momentum through Kepler's and Newton's entirely deterministic laws.

References

Heisenberg W. "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanick" (translated title: "On the illustrative content of quantum-mechanical kinetic and mechanic") *Zeitschrift für Physik* **43**, 172 (1927)

Werner Heisenberg (1901–1976)
Established uncertainty principle

6

The Schrödinger equation

6.1 The time-dependent Schrödinger equation

The Schrödinger equation is the key equation of quantum mechanics (Schrödinger 1925, 1926a, 1926b). This second order, partial differential equation determines the spatial shape and the temporal evolvement of a wave function in a given potential and for given boundary conditions. The one-dimensional Schrödinger equation is used when the particle of interest is confined to one spatial dimension, for example the x axis. Here, we restrict our considerations to such one-dimensional cases. Due to the one-dimensional nature of many semiconductor heterostructures, the one-dimensional Schrödinger equation is sufficient for most applications. To derive the one-dimensional Schrödinger equation, we start with the total-energy equation, *i. e*. the sum of kinetic and potential energy

$$\frac{p^2}{2\,m} + U(x) \;=\; E_{\text{total}} \;. \tag{6.1}$$

Substitution of the dynamical variables by their quantum mechanical operators which act on the wave function $\Psi(x, t)$ yields the ***one-dimensional time-dependent Schrödinger equation***

$$\boxed{-\frac{\hbar^2}{2\,m}\frac{\partial^2}{\partial x^2}\,\Psi(x,t) + U(x)\,\Psi(x,t) \;=\; -\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial t}\,\Psi(x,t)} \tag{6.2}$$

The left side of this equation can be rewritten by using the Hamilton or total-energy operator

$$H \;=\; -\frac{\hbar^2}{2\,m}\frac{\partial^2}{\partial x^2} + U(x) \;. \tag{6.3}$$

Using the notation of the Hamilton operator, the time-dependent Schrödinger equation can be written as

$$\boxed{H\,\Psi(x,t) \;=\; -\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial t}\,\Psi(x,t)} \tag{6.4}$$

Since the Schrödinger equation is a *partial differential equation*, the ***product method*** can be used to separate the equation into a spatial and a temporal part

$$\boxed{\Psi(x,t) \;=\; \psi(x)\;\mathrm{f}(t)} \tag{6.5}$$

where $\psi(x)$ depends only on x and $\mathrm{f}(t)$ depends only on t. Insertion of Eq. (6.5) into the Schrödinger equation yields

$$\frac{1}{\psi(x)} H\,\psi(x) \;=\; \frac{\mathrm{i}\,\hbar}{\mathrm{f}(t)}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{f}(t)\,. \tag{6.6}$$

The left side of this equation depends on x only, while the right side depends only on t. Because x and t are completely independent variables, the equation can be true, only if both sides are constant.

$$\frac{\mathrm{i}\,\hbar}{\mathrm{f}(t)}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{f}(t) \;=\; \text{const.} \tag{6.7}$$

Tentatively this constant is designated as const. = E where the meaning of E will become evident below. Integration of Eq. (6.7) yields

$$\mathrm{f}(t) \;=\; \mathrm{e}^{-\mathrm{i}\,(E/\hbar)t}\,. \tag{6.8}$$

By using $\mathrm{e}^{-\mathrm{i}(E/\hbar)t} = \mathrm{e}^{-\mathrm{i}\omega t}$, we can identify the angular frequency of oscillation of the quantum particle as $\omega = E/\hbar$, or, $E = \hbar\omega$ (Planck's relation). Insertion of this result into Eq. (6.5) yields the time-dependent wave function

$$\boxed{\Psi(x,t) \;=\; \psi(x)\,\mathrm{e}^{-\mathrm{i}\,(E/\hbar)t}}\,. \tag{6.9}$$

Since E is real, then the wave function has an *amplitude* $\psi(x)$ and a *phase* $\exp(-\mathrm{i}Et/\hbar)$. The amplitude and phase representation is convenient for many applications. To find the physical meaning of the real quantity E, we calculate the expectation value of the total energy using the wave function obtained from the product method.

$$\left\langle E_{\text{total}} \right\rangle \;=\; \int_{-\infty}^{\infty} \psi^{*}(x)\,\mathrm{f}^{*}(t)\left(-\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial t}\right)\psi(x)\,\mathrm{f}(t)\,\mathrm{d}x \;=\; \mathrm{e}^{\frac{\mathrm{i}Et}{\hbar}}\,\mathrm{e}^{-\frac{\mathrm{i}Et}{\hbar}}\,E\int_{-\infty}^{\infty}\psi^{*}(x)\,\psi(x)\,\mathrm{d}x \;=\; E\,. \tag{6.10}$$

Because the wave function is normalized, that is $\langle\psi(x)|\,\psi(x)\rangle = 1$, the constant designated as *E is the expectation value of the total energy*.

6.2 The time-independent Schrödinger equation

The time-independent Schrödinger equation is obtained by inserting the wave function obtained from the product method, Eq. (6.9) into the time-dependent Schrödinger equation (see Eq. 6.2). One obtains

$$\boxed{-\frac{\hbar^2}{2\,m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\psi(x) \;+\; U(x)\,\psi(x) \;=\; E\,\psi(x)} \tag{6.11}$$

which is the ***time-independent Schrödinger equation***. Using the hamiltonian operator, one obtains

$$H\,\psi(x) \;=\; E\,\psi(x)\,. \tag{6.12}$$

Since H is an operator and E is a real number, the Schrödinger equation has the ***form of an eigenvalue equation***. The eigenfunctions $\psi_n(x)$ and the eigenvalues E_n are found by solving the Schrödinger equation.

The eigenvalues of the Schrödinger equation, E_n, are *discrete*, that is only certain energy values are allowed, all other energies are disallowed or forbidden. The energy eigenvalues are also called ***eigenenergies*** or ***eigenstate energies***. The lowest eigenstate energy is the ***ground state energy***. All higher energies are called of ***excited state energies***.

The solution of the Schrödinger equation and the eigenstate energies and wave functions of a physical system are of great importance, because the knowledge of $\psi_n(x)$ and E_n implies the knowledge of *all* relevant physical parameters. It is the purpose of the next sections to get familiar with the properties of the Schrödinger equation and its solutions.

6.3 The superposition principle

Mathematically speaking the Schrödinger equation is a *linear*, second order, partial differential equation. Any linear differential equation allows for the superposition of its solutions. That is, if Ψ_n and Ψ_m are solutions of the Schrödinger equation, then any *linear combination* of Ψ_n and Ψ_m are solutions as well. That is, a new solution of Schrödinger's equation is given by

$$\Psi(x,t) \;=\; A\,\Psi_n(x,t) \;+\; B\,\Psi_m(x,t) \tag{6.13}$$

where, A and B are real constants. For practical physical problems, the Schrödinger equation has always more than one solution. Thus, the superposition principle can be applied to all physical problems in order to obtain a new solution. The new solution $\Psi(x, t)$ must be normalized as well, that is $\langle \Psi \mid \Psi \rangle = 1$.

6.4 The orthogonality of eigenfunctions

If two eigenfunctions $\psi_n = \psi_n(x)$ and $\psi_m = \psi_m(x)$ are solutions of the Schrödinger equation and the two eigenfunctions belong to different energies E_n and E_m (so that $E_n \neq E_m$), then the eigenfunctions are ***orthogonal***:

$$\left\langle \psi_n \middle| \psi_m \right\rangle \;=\; 0\,. \tag{6.14}$$

This equation can be proven by starting with the Schrödinger equation for $\psi_n(x)$ and the complex conjugate equation for ψ_m, that is

$$-\frac{\hbar^2}{2\,m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi_n \;+\; U\,\psi_n \;=\; E_n\,\psi_n \tag{6.15}$$

and

$$-\frac{\hbar^2}{2\,m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi_m^* \;+\; U\,\psi_m^* \;=\; E_m\,\psi_m^*\,. \tag{6.16}$$

Pre-multiplication of Eq. (6.15) with ψ_m* and of Eq. (6.16) with ψ_n and subtraction of the two resulting equations yields

$$\frac{\hbar^2}{2\,m}\left(\psi_n\,\frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\psi_m^*\;-\;\psi_m^*\,\frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\psi_n\right)\;=\;\left(E_n\;-\;E_m\right)\psi_n\,\psi_m^*. \tag{6.17}$$

Using the identity

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\psi_n\,\frac{\mathrm{d}}{\mathrm{d}x}\,\psi_m^*\;-\;\psi_m^*\,\frac{\mathrm{d}}{\mathrm{d}x}\,\psi_n\right)\;=\;\psi_n\,\frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\psi_m^*\;-\;\psi_m^*\,\frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\psi_n \tag{6.18}$$

and integrating over x yields

$$\frac{\hbar^2}{2\,m}\int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}x}\left(\psi_n\,\frac{\mathrm{d}}{\mathrm{d}x}\,\psi_m^*\;-\;\psi_m^*\,\frac{\mathrm{d}}{\mathrm{d}x}\,\psi_n\right)\mathrm{d}x\;=\;\left(E_n\;-\;E_m\right)\int_{-\infty}^{\infty}\psi_m^*\,\psi_n\,\mathrm{d}x. \tag{6.19}$$

The integral on the left side of the equation simplifies to

$$\int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}x}\left(\psi_n\,\frac{\mathrm{d}}{\mathrm{d}x}\,\psi_m^*\;-\;\psi_m^*\,\frac{\mathrm{d}}{\mathrm{d}x}\,\psi_n\right)\mathrm{d}x\;=\;\left[\psi_n\,\frac{\mathrm{d}}{\mathrm{d}x}\,\psi_m^*\;-\;\psi_m^*\,\frac{\mathrm{d}}{\mathrm{d}x}\,\psi_n\right]_{-\infty}^{+\infty} \tag{6.20}$$

This expression is zero due to the normalization condition which requires that $\psi\,(x \to \pm\,\infty) = 0$. Hence, we obtain the following condition for the eigenvalues and eigenfunctions

$$\frac{2\,m}{\hbar^2}\left(E_n\;-\;E_m\right)\int_{-\infty}^{\infty}\psi_m^*\,\psi_n\,\mathrm{d}x\;=\;0\;. \tag{6.21}$$

Because $E_n \neq E_m$, this equation can be true, only if

$$\int_{-\infty}^{\infty}\psi_m^*\,\psi_n\,\mathrm{d}x\;=\;0 \qquad\qquad (m \neq n) \tag{6.22}$$

which concludes the proof that ψ_m and ψ_n are orthogonal. Together with the normalization condition, one obtains

$$\int_{-\infty}^{\infty}\psi_m^*\,\psi_n\,\mathrm{d}x\;=\;\begin{cases}0 & (m\;\neq\;n)\\ 1 & (m\;=\;n)\,.\end{cases} \tag{6.23}$$

This result can be also written as

$$\left\langle\psi_m\middle|\,\psi_n\right\rangle\;=\;\delta_{mn} \tag{6.24}$$

where δ_{mn} is the ***Kronecker delta*** which is defined as

$$\delta_{mn}\;=\;\begin{cases}0 & (m\;\neq\;n)\\ 1 & (m\;=\;n)\,.\end{cases} \tag{6.25}$$

6.5 The complete set of eigenfunctions

Consider a practical physical problem given by the potential energy $U(x, y, z)$. Assume further that at least one solution of the Schrödinger equation exists for the potential energy $U(x, y, z)$. Generally, the number of solutions is large but finite. The solutions of the Schrödinger equation are designated a set of solutions. Such a set of solutions is a *complete set of solutions*, if it contains all possible solutions. If, in addition, each solution of the set is normalized and if the solutions are orthogonal, then the solutions are called an orthogonal, normal, complete set or, abbreviated, a ***orthonormal complete set*** of solutions. Such an orthonormal complete set of solutions provides *any* solution of a physical problem by superposition (linear combination) of the individual solutions.

Exercise 1: ***Solutions of the Schrödinger equation in a constant potential***. Assume that the potential $U(x)$ is a *constant*. What are the mathematical functions that are possible solutions of the one-dimensional Schrödinger equation in such a potential?

Solution: For $U(x)$ = constant, the Schrödinger equation has the form $(d^2/dx^2)\ \psi(x) \propto \pm\psi(x)$. This equation has either an exponential or a sinusoidal solution. An example is shown in ***Fig***. 6.1.

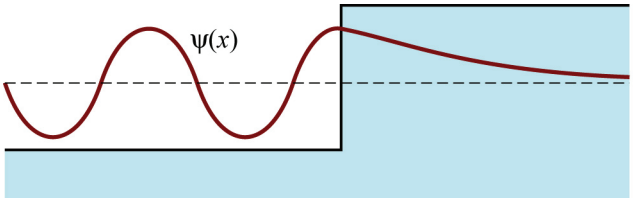


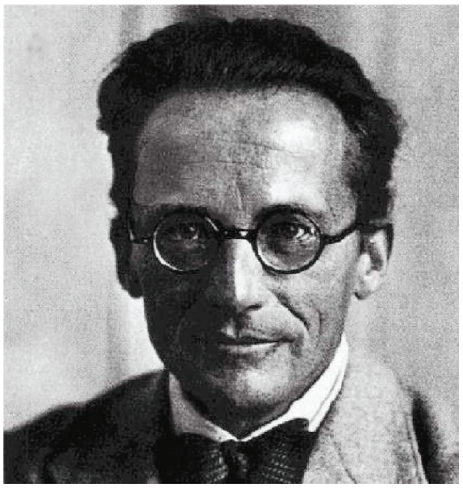
Fig. 6.1. Sinusoidal (left) and exponential (right) solution of the Schrödinger equation in a constant potential.

References

Schrödinger E. "Quantsierung als Eigenwertproblem" (translated title: "Quantization as an eigenvalue problem") (Part 1) *Annalen der Physik* **79**, 361 (1925)

Schrödinger E. "Quantisierung als Eigenwertproblem (translated title: "Quantification as an eigenvalue problem)" *Annalen der Physik* **80,** 437 (1926a)

Schrödinger E. "Quantisierung als Eigenwertproblem" (translated title: "Quantization as an eigenvalue problem") (Part 4) *Annalen der Physik* **81**, 109 (1926b)



Erwin Schrödinger (1887–1961)
Established quantum mechanical wave equation

7

Applications of the Schrödinger equation in nonperiodic structures

7.1 Electron in a constant potential

The wave function of an electron in a piecewise constant potential is shown in ***Fig***. 7.1. The potential is given by

$$U_{\mathrm{I}}(x) = U_{\mathrm{I}} \qquad x < L \tag{7.1}$$

$$U_{\mathrm{II}}(x) = U_{\mathrm{II}} \qquad x > L. \tag{7.2}$$

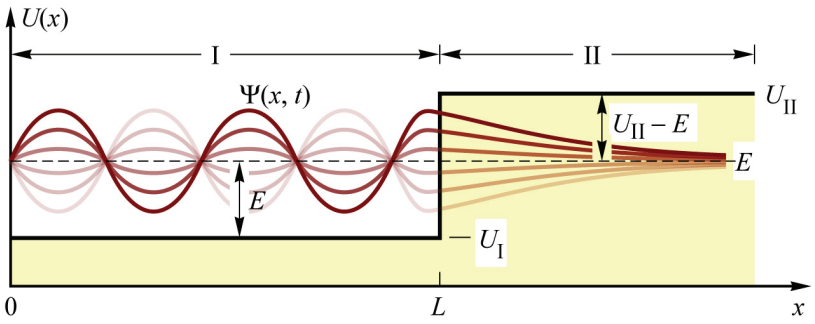


Fig. 7.1. Electron wave function in a constant potential.

Given the fact that the eigenfunctions of the operator d^2/dx^2 (i.e. the x-dependent part of the Hamiltonian operator in a constant potential) are either sinusoidal or exponentially decreasing (or increasing) functions, allows us to write the solutions of the time-dependent Schrödinger equation. Assuming that the energy of the electron is $U_{\mathrm{I}} < E < U_{\mathrm{II}}$, the solution of the Schrödinger equation in ***Region I*** is given by

$$\Psi(x,t) = A\, \mathrm{e}^{\mathrm{i}(kx-\omega t)} \qquad \text{for } x < L \tag{7.3}$$

where A is a normalization constant,

$$k = \sqrt{2\,m(E-U_{\mathrm{I}})\,/\,\hbar^2}\ , \tag{7.4}$$

and

$$\omega = E\,/\,\hbar\ . \tag{7.5}$$

The solution of the Schrödinger equation in ***Region II*** is given by

$$\Psi(x,t) = A\, \mathrm{e}^{-\kappa x}\, \mathrm{e}^{\mathrm{i}\omega t} \qquad \text{for } x > L \tag{7.6}$$

where

$$\kappa \;=\; \sqrt{2\,m\,(U_{\mathrm{II}} - E)\,/\,\hbar^2}\quad . \tag{7.7}$$

Insertion of these solutions into the Schrödinger equation allows one to verify that they indeed are correct solutions of the Schrödinger equation.

The time-dependent oscillatory factor of the wave function (i.e. exp iωt) always appears in this form. Therefore, we will not be concerned with the time-dependent factor in our subsequent discussions.

7.2 The infinite square-shaped quantum well

The infinite square-shaped well potential is the simplest of all possible potential wells. The infinite square well potential is illustrated in ***Fig***. 7.2(a) and is defined as

$$U(x) \;=\; 0 \qquad \left(-\tfrac{1}{2}L \;\le x\; \le \tfrac{1}{2}L\right) \tag{7.8}$$

$$U(x) \;=\; \infty \qquad \left(|x| \;>\; \tfrac{1}{2}L\right). \tag{7.9}$$

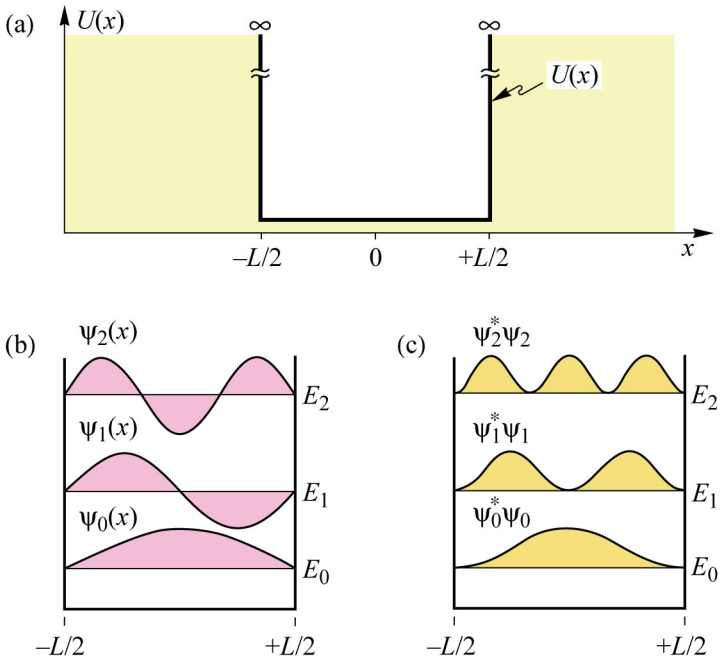


Fig. 7.2. (a) Schematic illustration of the infinite square well potential. The solutions of this potential well are shown in terms of (b) eigenfunctions $\psi_n(x)$, (b) eigenstate energies E_n, and (c) probability densities $\psi_n^*\psi_n$.

To find the stationary solutions for $\psi_n(x)$ and E_n we must find functions for $\psi_n(x)$, which satisfy the Schrödinger equation. The time-independent Schrödinger equation contains only the differential operator $\mathrm{d}/\mathrm{d}x$, whose eigenfunctions are exponential or sinusoidal functions. Since the Schrödinger equation has the form of an eigenvalue equation, it is reasonable to try only eigenfunctions of the differential operator. Furthermore, we assume that $\psi_n(x) = 0$ for $|\,x\,| > L/2$, because the potential energy is infinitely high in the barrier regions. Since the 3rd Postulate requires that the wave function be continuous, the wave function must have zero amplitude at the two potential discontinuities, that is $\psi_n(x = \pm L/2) = 0$. We therefore employ sinusoidal functions and differentiate between states of even and odd symmetry. We write for even-

symmetry states

$$\psi_n(x) = A \cos \frac{(n+1)\,\pi\, x}{L} \qquad \left(n = 0, 2, 4 \cdots \quad \text{and} \quad |x| \le \frac{L}{2} \right) \tag{7.10}$$

and for odd-symmetry states

$$\psi_n(x) = A \sin \frac{(n+1)\,\pi\, x}{L} \qquad \left(n = 1, 3, 5 \cdots \quad \text{and} \quad |x| \le \frac{L}{2} \right). \tag{7.11}$$

Both functions have a finite amplitude in the well-region ($|x| \le L/2$) and they have zero amplitude in the barriers, that is

$$\psi_n(x) = 0 \qquad \left(n = 0, 1, 2 \cdots \quad \text{and} \quad |x| > \frac{L}{2} \right) \tag{7.12}$$

The shapes of the three lowest wave functions (n = 0, 1, 2 …) are shown in ***Fig***. 7.2(b). In order to normalize the wave functions, the constant A must be determined. The condition $\langle \psi \mid \psi \rangle = 1$ yields

$$A = \sqrt{2/L}\ . \tag{7.13}$$

One can verify that Eqs. (7.10) and (7.11) are solutions of the infinite square well by inserting the normalized wave functions into the Schrödinger equation. Insertion of the ground-state wave function (n = 0) into the Schrödinger equation yields

$$-\frac{\hbar^2}{2\,m} \frac{\mathrm{d}^2}{\mathrm{d}x^2} \sqrt{\frac{2}{L}}\, \cos\left(\frac{\pi\, x}{L} \right) = E_0 \sqrt{\frac{2}{L}}\, \cos\left(\frac{\pi\, x}{L} \right). \tag{7.14}$$

Calculating the derivative on the left-hand side of the equation yields the ground state energy of the infinite square well

$$E_0 = \frac{\hbar^2}{2\,m} \left(\frac{\pi}{L} \right)^2. \tag{7.15}$$

The excited state energies (n = 1, 2, 3 …) can be evaluated analogously. One obtains the eigenstate energies in the infinite square well as

$$E_n = \frac{\hbar^2}{2\,m} \left[\frac{(n+1)\,\pi}{L} \right]^2 \qquad (n = 0, 1, 2 \ldots). \tag{7.16}$$

The spacing between two adjacent energy levels, that is $E_n - E_{n-1}$, is proportional to n. Thus, the energetic spacing between states *increases* with energy. The energy levels are schematically shown in ***Fig***. 7.2(b) for the infinite square well.

The probability density of a particle described by the wave function ψ is given by $\psi^* \psi$ (2nd Postulate). The probability densities of the three lowest states are shown in ***Fig***. 7.2(c).

The eigenstate energies are, as already mentioned, *expectation values of the total energy* of the respective state. It is therefore interesting to know if the eigenstate energies are purely kinetic, purely potential, or a mixture of both. The expectation value of the kinetic energy of the ground state is calculated according to the 5th Postulate:

$$\left\langle E_{\mathrm{kin},0} \right\rangle = \left\langle \psi_0 \left| \frac{p^2}{2\,m} \right| \psi_0 \right\rangle . \tag{7.17}$$

Using the momentum operator $p = (\hbar / \mathrm{i})\ (\mathrm{d} / \mathrm{d}x)$ one obtains the expectation value of the kinetic energy of the ground state

$$\left\langle E_{\mathrm{kin},0} \right\rangle = \frac{\hbar^2}{2\,m} \left(\frac{\pi}{L} \right)^2 \tag{7.18}$$

which is identical to the total energy given in Eq. (7.15). Evaluation of kinetic energies of all other states yields

$$\left\langle E_{\mathrm{kin},n} \right\rangle = \frac{\hbar^2}{2\,m} \left[\frac{(n+1)\,\pi}{L} \right]^2 . \tag{7.19}$$

The kinetic energy coincides with the total energy given in Eq. (7.16). Thus, the energy of a particle in an infinite square well is purely kinetic. The particle has no potential energy.

Second method: Matching the de Broglie wavelength to the width of quantum well. We next turn to a second, more intuitive method to obtain the wave functions of the infinite potential well. This second method is based on the de Broglie wave concept. Recall that the de Broglie wave is defined for a constant momentum p, that is, for a particle in a *constant* potential. The energy of the wave is purely kinetic. In order to find solutions of the infinite square well, we match the de Broglie wavelength to the width of the quantum well according to the condition

$$\frac{1}{2}\,\lambda\,(n+1) = L \qquad (n = 0,\ 1,\ 2\ \ldots) \tag{7.20}$$

In this equation, multiples of half of the de Broglie wavelength are matched to the width of the quantum well. Expressing the kinetic energy in terms of the de Broglie wavelength, that is

$$E = \frac{p^2}{2\,m} = \frac{1}{2\,m} \left(\frac{2\,\pi\,\hbar}{\lambda} \right)^2 \tag{7.21}$$

and inserting Eq. (7.20) into Eq. (7.21) yields

$$E_n = \frac{\hbar^2}{2\,m} \left[\frac{(n+1)\,\pi}{L} \right]^2 \qquad (n = 0,\ 1,\ 2\ \ldots) . \tag{7.22}$$

This equation is identical to Eq. (7.19) which was obtained by the solution of the Schrödinger equation. The de Broglie wave concept yields the correct solution of the infinite potential well,

because (*i*) the particle is confined to the constant potential of the well region, (*ii*) the energy of the particle is purely kinetic, and (*iii*) the wave function is sinusoidal.

The infinite square shaped quantum well is the simplest of all potential wells. The wave functions (eigenfunctions) and energies (eigenvalues) in an infinite square well are relatively simple. There is a large number of potential wells with other shapes, for example the square well with *finite* barriers, parabolic well, triangular well, or V-shaped well. The exact solutions of these wells are more complicated. Several methods have been developed to calculate approximate solutions for arbitrary shaped potential wells. These methods will be discussed in the chapter on quantum wells in this book.

7.3 The asymmetric and symmetric finite square-shaped quantum well

In contrast to the infinite square well, the *finite square well* has barriers of finite height. The potential of a finite square well is shown in ***Fig***. 7.3. The two barriers of the well have a different height and therefore, the structure is denoted *asymmetric* square well. The potential energy is constant within the three regions I, II, and III, as shown in ***Fig***. 7.3. In order to obtain the solutions to the Schrödinger equation for the square well potential, the solutions in a *constant potential* will be considered first.

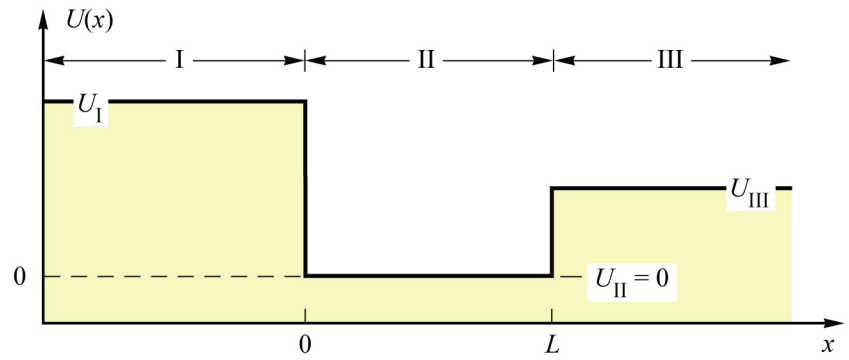


Fig. 7.3. Asymmetric square well potential with well width L and barrier heights U_I and U_{III}.

Assume that a particle with energy E is in a constant potential U. Then two cases can be distinguished, namely $E > U$ and $E < U$. In the ***first case*** ($E > U$) the general solution to the time-independent one-dimensional Schrödinger equation is given by

$$\psi(x) = A\cos kx + B\sin kx \tag{7.23}$$

where A and B are constants and

$$k = \sqrt{2mE/\hbar^2} \quad . \tag{7.24}$$

Insertion of the solution into the Schrödinger equation proves that it is indeed a correct solution. Thus the wave function is an *oscillatory sinusoidal function* in a constant potential with $E > U$. In the ***second case*** ($E < U$), the solution of the time-independent one-dimensional Schrödinger equation is given by

$$\psi(x) = C\,e^{\kappa x} + D\,e^{-\kappa x} \tag{7.25}$$

where C and D are constants and

$$\kappa = \sqrt{\frac{2m(U-E)}{\hbar^2}} = \sqrt{\frac{2mU}{\hbar^2} - k^2} \quad . \tag{7.26}$$

Again, the insertion of the solution into the Schrödinger equation proves that it is indeed a correct solution. Thus the wave function is an *exponentially growing* or *decaying* function in a constant potential with $E < U$.

Next, the solutions of an asymmetric and symmetric square well will be calculated. The potential energy of the well is piecewise constant, as shown in ***Fig***. 7.3. Having shown that the wave functions in a constant potential are either sinusoidal or exponential, the wave functions in the three regions I ($x \le 0$), II ($0 < x < L$), and III ($x \ge L$), can be written as

$$\psi_I(x) = A\, e^{\kappa_I x} \tag{7.27}$$

$$\psi_{II}(x) = A \cos kx + B \sin kx \tag{7.28}$$

$$\psi_{III}(x) = (A \cos kL + B \sin kL)\, e^{-\kappa_{III}(x-L)} \tag{7.29}$$

where A and B are unknown normalization constants. In this solution, the first boundary condition of the 3rd Postulate, *i. e.* $\psi_I(0) = \psi_{II}(0)$ and $\psi_{II}(L) = \psi_{III}(L)$, is already satisfied. From the second boundary condition of the 3rd Postulate, *i. e.* $\psi_I'(0) = \psi_{II}'(0)$ and $\psi_{II}'(L) = \psi_{III}'(L)$, the following two equations are obtained

$$A\kappa_I - Bk = 0 \tag{7.30}$$

$$A(\kappa_{III} \cos kL - k \sin kL) + B(\kappa_{III} \sin kL + k \cos kL) = 0 \; . \tag{7.31}$$

This homogeneous system of equations has solutions, only if the determinant of the system vanishes. From this condition, one obtains

$$\boxed{\tan kL = \frac{kL(\kappa_I L + \kappa_{III} L)}{k^2 L^2 - \kappa_I L \kappa_{III} L}} \tag{7.32}$$

which is the ***eigenvalue equation*** of the *finite asymmetric square well*.

For the *finite symmetric square well*, which is of great practical relevance, the ***eigenvalue equation*** is given by

$$\boxed{\tan kL = \frac{2kL\,\kappa L}{k^2 L^2 - \kappa^2 L^2}} \tag{7.33}$$

where $\kappa = \kappa_I = \kappa_{III}$. If κ is expressed as a function of k (see Eq. 7.26), then Eq. (7.33) depends only on a single variable, *i. e.*, k. Solving the eigenvalue equation yields the eigenvalues of k and, by using Eqs. (7.24) and (7.26), the allowed energies E and decay constants κ, respectively. The allowed energies are also called the ***eigenstate energies*** of the potential.

Inspection of Eq. (7.33) yields that the eigenvalue equation has a ***trivial solution*** $kL = 0$ (and thus $E = 0$) which possesses no practical relevance. Non-trivial solutions of the eigenvalue equation can be obtained by a graphical method. ***Figure*** 7.4 shows the graph of the left-hand and

right-hand side of the eigenvalue equation. The dashed curve represents the right-hand side of the eigenvalue equation. The intersections of the dashed curve with the periodic tangent function are the solutions of the eigenvalue equation. The quantum state with the lowest non-trivial solution is called the ***ground state*** of the well. States of higher energy are referred to as ***excited states***.

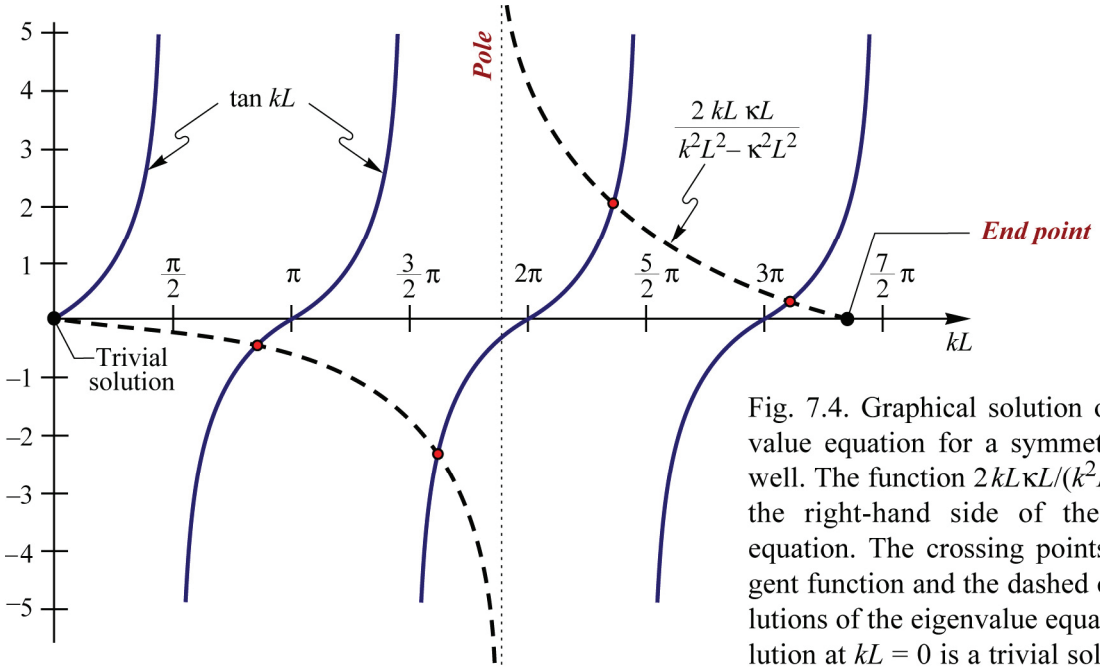


Fig. 7.4. Graphical solution of the eigenvalue equation for a symmetric quantum well. The function $2kL\kappa L/(k^2L^2 - \kappa^2L^2)$ is the right-hand side of the eigenvalue equation. The crossing points of the tangent function and the dashed curve are solutions of the eigenvalue equation. The solution at $kL = 0$ is a trivial solution having no practical relevance.

The dashed curve shown in ***Fig***. 7.4 has two significant points, namely a *pole* and an *end point*. The dashed curve has a ***pole*** when the denominator of the right-hand side of the eigenvalue equation vanishes, *i. e*., when $kL = \kappa L$. Using Eq. (7.26), it is given by

$$\textbf{\textit{Pole}}: \qquad k\,L\big|_{\text{Pole}} = \sqrt{m\,U\,/\,\hbar^2}\;\;L \tag{7.34}$$

The dashed curve *ends* when $k = (2mU/\hbar^2)^{1/2}$. If k exceeds this value, the square root in Eq. (7.26) becomes imaginary. The ***end point*** of the dashed curve is thus given by

$$\textbf{\textit{End point}}: \qquad k\,L\big|_{\text{End point}} = \sqrt{2\,m\,U\,/\,\hbar^2}\;\;L \tag{7.35}$$

There are no further bound state solutions to the eigenvalue equation beyond the end point.

Now that the eigenvalues of k and κ are known, they are inserted into Eqs. (7.30) and (7.31); this allows for the determination of the constants A and B and the wave functions. Thus the allowed energies and the wave functions of the square well have been determined.

It is possible to show that all states with even quantum numbers ($n = 0, 1, 2 \ldots$) are of even symmetry with respect to the center of the well, *i. e*. $\psi(x) = \psi(-x)$. All states with odd quantum numbers ($n = 1, 3, 5 \ldots$) are of odd symmetry with respect to the center of the well, *i. e*. $\psi(x) = -\psi(-x)$. The even and odd state wave functions in the well are thus of the form

$$\psi_{II}(x) = A_{II} \cos\left[k_n \left(x - \frac{L}{2} \right) \right] \qquad \text{(for } n = 0, 2, 4 \ldots\text{)} \tag{7.36}$$

and

$$\psi_{II}(x) = A_{II} \sin\left[k_n \left(x - \frac{L}{2} \right) \right] \qquad \text{(for } n = 1, 3, 5 \ldots\text{)} . \tag{7.37}$$

The proof of these equations is left to the reader. The three lowest wave functions of a symmetric square well are shown in ***Fig***. 7.5.

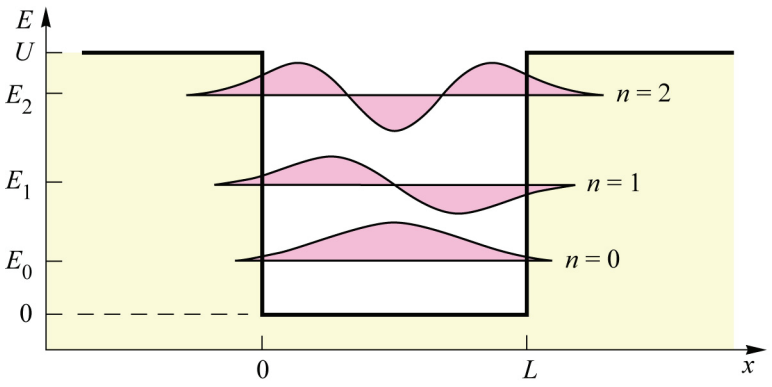


Fig. 7.5. Schematic illustration of the three lowest wave functions of the symmetric quantum well.

Exercise 1: ***Boundary condition in semiconductor quantum wells***. The boundary conditions for the wave function at an interface between two media I and II are given by $\psi_I(x) = \psi_{II}(x)$ and $\psi_I'(x) = \psi_{II}'(x)$ as stated in the 3rd Postulate. These boundary conditions apply to situations common in particle physics, in which the particle mass m *does not change* when going from a Medium I across a boundary to a Medium II. However, the effective mass (m^*) *changes* as electrons transfer from one semiconductor to another. This change in effective mass requires a modification of the second boundary condition and the ***modified second boundary condition*** is given by

$$\boxed{\frac{1}{m_I^*} \frac{d\psi_I(x)}{dx} = \frac{1}{m_{II}^*} \frac{d\psi_{II}(x)}{dx}} \tag{7.38}$$

The first boundary condition, namely $\psi_I(x) = \psi_{II}(x)$, is still valid and this condition does not need to be modified.

The modified second boundary condition can be derived from the requirement of a *constant current density* at the boundary. The current density of an electron moving with constant velocity v across an interface is given by $J = eV^{-1} v$, where V is the unit volume. Expressing the current density in terms of the electron momentum yields $J = eV^{-1} p / m^*$. Using the 4th Postulate, a corresponding quantum mechanical expression is found, *i. e.*

Classical	***Quantum mechanical***
$e V^{-1} \frac{p}{m^*}$	$e V^{-1} \frac{1}{m^*} \frac{\hbar}{\mathrm{i}} \frac{d\psi(x)}{dx}$

The quantum mechanical expression elucidates that the current density at an interface is constant, only if $(m^*)^{-1}\,[\mathrm{d}\psi(x)/\mathrm{d}x]$ is constant across the interface.

A rigorous derivation of the quantum mechanical current density was given by Flügge (1971). It is given by

$$J = \frac{e\hbar}{2\,m\,\mathrm{i}}\left(\psi^*\,\nabla\psi - \psi\,\nabla\psi^*\right) \tag{7.39}$$

Bastard (1981) first showed that the second boundary condition must be modified in semiconductor heterostructures according to Eq. (7.38).

Apply the modified boundary condition to an asymmetric semiconductor quantum well structure and derive the eigenvalue equation. Assume that the effective masses are m_I^*, m_II^*, and m_III^* in the first barrier, well, and second barrier region, respectively.

Result:
$$\tan k\,L = \frac{\dfrac{k\,L}{m_\mathrm{II}^*}\left(\dfrac{\kappa_\mathrm{I}\,L}{m_\mathrm{I}^*} + \dfrac{\kappa_\mathrm{III}\,L}{m_\mathrm{III}^*}\right)}{\dfrac{k^2L^2}{m_\mathrm{II}^{*2}} - \dfrac{\kappa_\mathrm{I}\,L}{m_\mathrm{I}^*}\,\dfrac{\kappa_\mathrm{III}\,L}{m_\mathrm{III}^*}} \tag{7.40}$$

What is the eigenvalue equation for the symmetric semiconductor quantum well with $\kappa = \kappa_\mathrm{I} = \kappa_\mathrm{III}$ and $m_\mathrm{I}^* = m_\mathrm{III}^*$?

Result:
$$\tan k\,L = \frac{2\,\dfrac{k\,L}{m_\mathrm{II}^*}\,\dfrac{\kappa\,L}{m_\mathrm{I}^*}}{\dfrac{k^2L^2}{m_\mathrm{II}^{*2}} - \dfrac{\kappa^2L^2}{m_\mathrm{I}^{*2}}} \tag{7.41}$$

What is the maximum value for *kL*, *i. e.*, *end point* of the dashed curve?

Result:
$$k\,L\big|_\mathrm{Endpoint} = \sqrt{2\,m_\mathrm{II}^*\,U\,/\,\hbar^2}\;L \tag{7.42}$$

Use the location of the end point to determine a condition for the quantum wells thickness under which a symmetric quantum well structure has *only* one bound state.

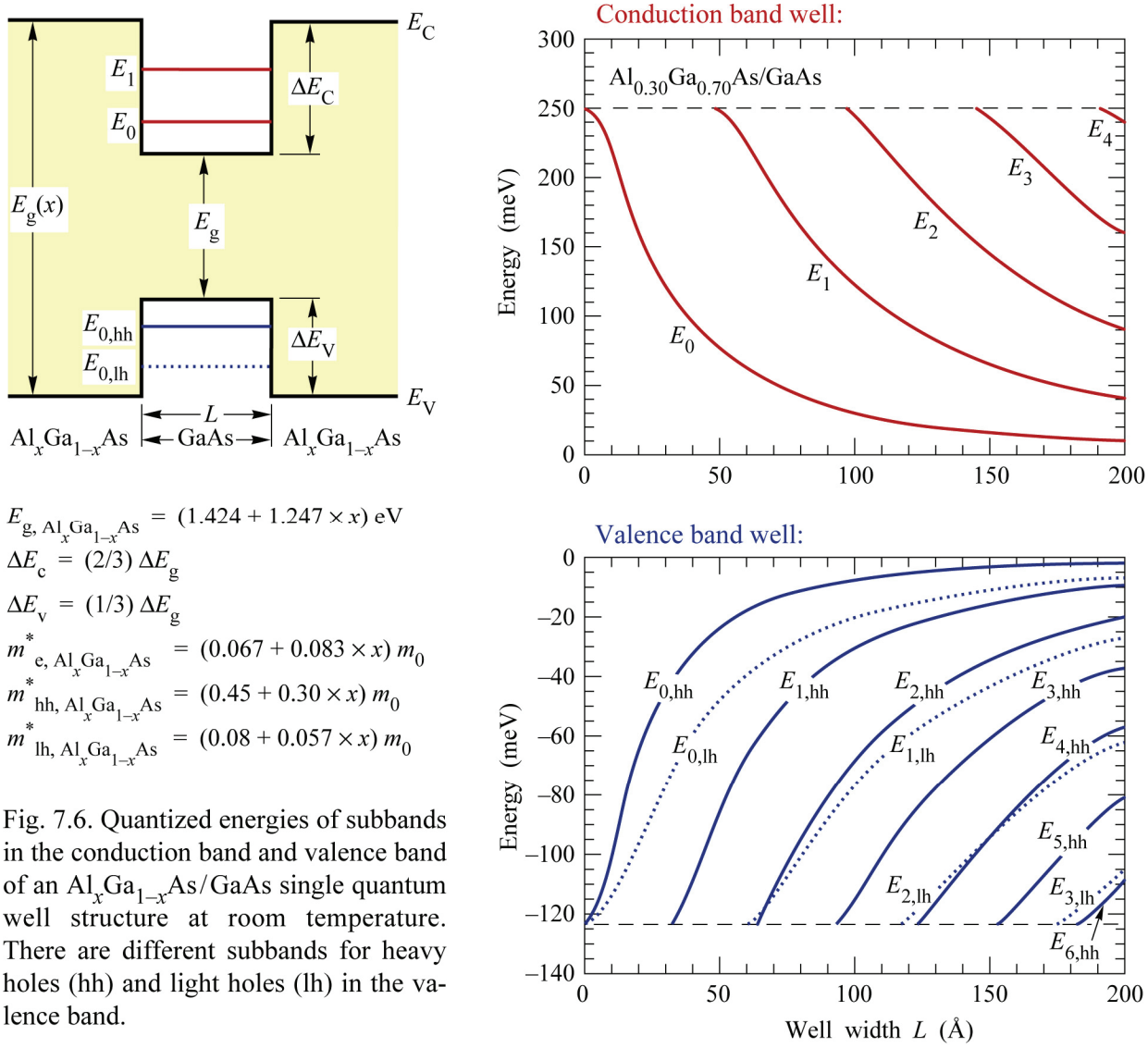
Result:
$$L < \frac{\pi\hbar}{\sqrt{2\,m_\mathrm{II}^*\,U}} \tag{7.43}$$

Is it possible for asymmetric or symmetric square well structures to have no bound states at all?

Answer: A symmetric square well always has at least one bound state. An asymmetric square well has no bound state for sufficiently small values of *L*.

Figure 7.6 shows the numerical solutions for bound states in the conduction band and valence band of an $Al_{0.30}Ga_{0.70}As\,/\,GaAs$ square-shaped quantum well. The graph reveals that there is

only one bound state in the conduction band well for well widths smaller than approximately 50 Å. Does this agree with the analytic result of Eq. (7.43)?



$$E_{g,\,Al_xGa_{1-x}As} = (1.424 + 1.247 \times x)\ \text{eV}$$
$$\Delta E_c = (2/3)\,\Delta E_g$$
$$\Delta E_v = (1/3)\,\Delta E_g$$
$$m^*_{e,\,Al_xGa_{1-x}As} = (0.067 + 0.083 \times x)\,m_0$$
$$m^*_{hh,\,Al_xGa_{1-x}As} = (0.45 + 0.30 \times x)\,m_0$$
$$m^*_{lh,\,Al_xGa_{1-x}As} = (0.08 + 0.057 \times x)\,m_0$$

Fig. 7.6. Quantized energies of subbands in the conduction band and valence band of an $Al_xGa_{1-x}As/GaAs$ single quantum well structure at room temperature. There are different subbands for heavy holes (hh) and light holes (lh) in the valence band.

References

Bastard G. "Superlattice band structure in the envelope-function approximation" *Physical Review B* **24**, 65693 (1981)

8

Applications of the Schrödinger equation in periodic structures

8.1 Free electrons

Before considering electrons in the periodic potential of a semiconductor crystal, we first consider electrons in free space that is in an environment in which no forces act on the particle. The lack of forces requires that the electrostatic potential, in which the electron propagates, is a constant. We first consider the propagation of the particle in the classical mechanics picture, then in the semi-classical picture, and finally in quantum-mechanical wave picture.

The propagation of a particle in the classical picture is described by newtonian mechanics. An electron with mass m and momentum $p = m\,v$ has the kinetic energy

$$E = \frac{1}{2} m v^2 = \frac{p^2}{2m} . \tag{8.1}$$

A particle with this energy will propagate at a constant velocity as long as no forces act on it.

A semi-classical description of the particle is obtained by taking into account the de Broglie relationship $p = \hbar\, k$. The free electron kinetic energy can then be written as

$$\boxed{E = \frac{\hbar^2\, k^2}{2\, m}} \tag{8.2}$$

This description includes the wave-like character of the particle by the wave vector $\vec{k}$ while it preserves the deterministic nature of the classical particle. Equation (8.2) is, therefore, a semi-classical representation of the propagation of the particle. Generally, the wave vector has components along the three axes of the cartesian coordinate system, that is $\vec{k} = (k_x, k_y, k_z)$. Hence $\vec{k}^{\,2} = k^2 = k_x^{\,2} + k_y^{\,2} + k_z^{\,2}$. Expressing $\vec{k}^{\,2}$ in Eq. (8.2) in terms of its x, y, and z components gives

$$E = \frac{\hbar^2}{2\, m} \left(k_x^2 + k_y^2 + k_z^2 \right). \tag{8.3}$$

In k space with the cartesian coordinates k_x, k_y, and k_z, Eq. (8.2) represents a sphere if the energy E is a constant. The sphere has a "radius" *in k-space* of $(2\,E\,m/\hbar^2)^{1/2}$. ***Thus the constant-energy surfaces in k-space of free electrons are spheres***. The shape of the constant-energy surfaces provides information about the propagation characteristics of carriers. A spherical constant-energy surface indicates that the propagation characteristics are *isotropic*.

The quantum-mechanical treatment of a free particle is based on the wave function $\Psi(x, y, z, t)$ which represents the particle. By introducing the wave function, the deterministic nature of classical mechanics is lost. The product method allows one to separate the wave

function into a time-dependent part and a time-independent part. The time-dependent part is given by the oscillatory term $\exp[-\mathrm{i}(E/\hbar)t]$. The time-independent part is denoted as $\psi(x, y, z)$. The complete wave function is given by the product of the two parts. Here, we are interested in the wave vector and the energy of the wave both of which do not depend on time for a freely propagating particle. Thus, the wave function $\psi(x, y, z)$ must satisfy the time-independent Schrödinger equation

$$-\frac{\hbar^2}{2m}\nabla^2\psi + U\psi = E\psi \tag{8.4}$$

where $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$ and $U = U(x, y, z)$ is the potential energy in the medium in which the particle propagates. Since we assumed that no forces act on the particle, the potential energy must be constant. For example, $U(x, y, z) = 0$. Assuming that the wave representing the particle propagates along the direction given by the vector $\vec{r}$, the solution of the Schrödinger equation is given by

$$\psi(\vec{r}) = A\,\mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{r}} \tag{8.5}$$

where the kinetic energy of the particle is given by $E = \hbar^2 k^2/(2m)$. Insertion of $\psi(\vec{r})$ into the time-independent Eq. (8.4) and using $\nabla^2 = \partial^2/\partial r^2$ yields that the kinetic energy of the wave as

$$E = \frac{\hbar^2 k^2}{2m}. \tag{8.6}$$

This equation is identical to Eq. (8.2) which was derived by using semi-classical arguments.

We have seen in this section that the propagation of a free particle is describable by the wave vector $\vec{k}$ in the semi-classical and in the quantum-mechanical picture. We have also seen that the surfaces of constant energy of a free particle are spheres in k space. In the next section, the propagation of electrons in the periodic lattice of a semiconductor will be discussed.

8.2 The Bloch theorem

We next consider the propagation of electrons in the periodic potential of a lattice. Consider an electron propagating along a straight line, for example along the direction of a vector $\vec{r}$ in the lattice. The electron is then exposed to periodic variations of the potential caused by the charged nuclei of the atoms forming the lattice and by the electrons of these atoms. The periodic potential is schematically shown in ***Fig***. 8.1. The potential energy of the lattice is periodic with a period $\vec{R}$. $\vec{R}$ is the vector of translational symmetry of the lattice as defined in Eq. (8.1). The periodicity of the potential energy in the lattice can be expressed by

$$U(\vec{r}) = U(\vec{r} + n\vec{R}) \qquad \text{for} \qquad n = 1, 2, 3 \ldots \tag{8.7}$$

To study the influence of this periodic lattice potential we must consider the influence of this potential on the quantum stationary states of the conduction electrons. In a constant potential (*i. e*. non-periodic potential), the wave function of electrons has the form $\exp(\mathrm{i}\,\vec{k}\cdot\vec{r})$, as shown in the proceeding section.

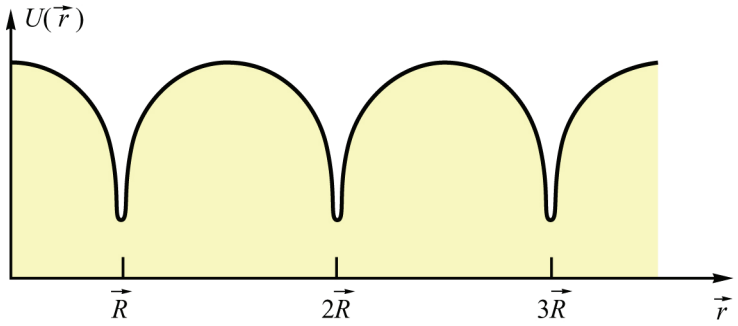


Fig. 8.1 Schematic illustration of a one-dimensional periodic potential caused by equally spaced atoms in a crystal lattice. The potential is periodic with period $\vec{R}$, that is $U(\vec{r}) = U(\vec{r} + \vec{R})$.

The propagation of electrons in a periodic potential was considered by Bloch (1928, 1930). He found that the wave function of an electron in a periodic lattice can be described by

$$\boxed{\psi_{nk}(\vec{r}) \;=\; u_{nk}\, e^{i\vec{k}\cdot\vec{r}}} \tag{8.8}$$

where u_{nk} has the same periodicity as the periodic potential, that is

$$u_{nk}(\vec{r}) \;=\; u_{nk}(\vec{r}+\vec{R}) \;=\; u_{nk}(\vec{r}+2\vec{R}) \;=\; \cdots \tag{8.9}$$

According to the ***Bloch theorem*** of Eq. (8.8), the wave function of an electron in a periodic potential consists of two parts, namely a lattice periodic part $u_{nk}(\vec{r})$, and the wave function of a free particle, that is exp (i $\vec{k}\cdot\vec{r}$). The product of both factors is the Bloch wave function, or short, the ***Bloch function***. The Bloch function plays an important role in solid state physics. Many theoretical models for example the Kronig–Penney model, are based on the Bloch wave function. It is important to note that the second factor of the Bloch function retains the same form as for free particles. That is, the lattice potential modulates the amplitude of the original free electron wave function through the function $u_{nk}(\vec{r})$. Using the periodicity of the function $u_{nk}(\vec{r})$, the Bloch theorem can be also expressed by

$$\psi_{nk}(\vec{r}+\vec{R}) \;=\; \psi_{nk}(r)\, e^{i\vec{k}\cdot\vec{R}} \tag{8.10}$$

where $\vec{R}$ is the period of the lattice.

We next have a closer look at the lattice-periodic part of the Bloch function, $u_{nk}(\vec{r})$. This function has the two subscripts n and k which indicate the dependence of the function on n and $\vec{k}$ (Note that the vector arrow is left off the subscript k). Generally the function depends on the electron wave vector $\vec{k}$ of the electron. The subscript n of the function is called the ***band index***. The function u_{nk} has a unique shape for the different energy bands of the semiconductor. For the conduction band, we will use the subscript c, *i. e.* u_{ck}. For the valence band, we write u_{vk}.

The Bloch theorem will not be proven here and the interested reader is referred to the literature (see, for example Ashcroft and Mermin, 1976). We will, however, analyze the Bloch function near $\vec{k} = 0$. The function $u_{nk}(\vec{r})$ is periodic with the period $R = |\vec{R}|$. It is therefore reasonable to assume that $u_{nk}(\vec{r})$ does not vary significantly for $|\vec{r}| << |\vec{R}|$. Consequently, the Bloch function will be dominated by the factor exp (i $\vec{k}\cdot\vec{r}$) if $|\vec{k}| << 2\pi/|\vec{R}|$. That is, in the regime of small values of k, the electron wave function is practically described by the free particle wave function. As a consequence, the ***dispersion relation*** of a particle in a periodic potential, $E(k)$, will be *parabolic* in this regime as well, similar to the parabolic dispersion

relation of free particles given in Eq. (8.6). For larger k values, the function $u_{nk}(\vec{r})$ cannot be neglected. Significant deviations from the parabolic dispersion are expected in this regime. Our expectation of a parabolic dispersion near $k = 0$ and significant deviations from the parabolic dispersion for larger k values will be confirmed in the next section.

8.3 The Kronig–Penney model

The electron wave function in a periodic potential is, as shown in the preceding section, given by the Bloch function. In this section, we apply the Bloch function to a very simple periodic potential, namely a one-dimensional square-shaped potential. The calculation will yield the allowed energy bands and the forbidden energy gaps as well as the E (k) relation for electrons, *i. e*. the relation between the energy E and the wave number k of the electron. The $E(k)$ relation is called the energy band structure or ***band structure*** of the lattice. The Schrödinger equation for a one-dimensional periodic potential was first solved by Kronig and Penney (1930). The calculation and the implications of this calculation are therefore referred to as the Kronig–Penney model.

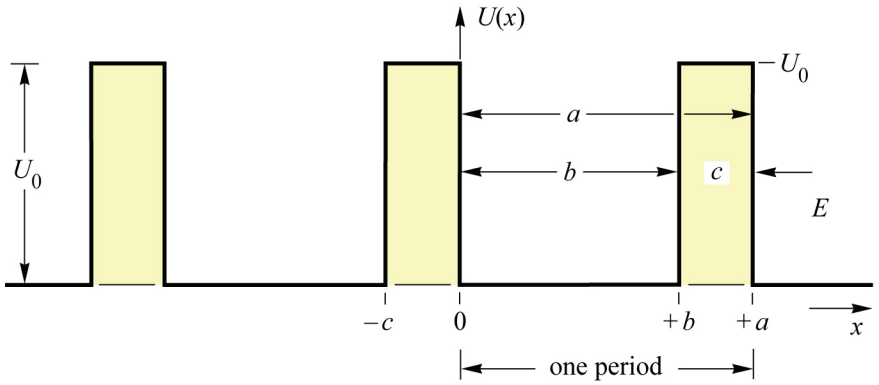


Fig. 8.2. Periodic square well potential used for the Kronig-Penney calculation. The height of the barriers is U_0 and the electron energy is denoted as E.

A simple one-dimensional potential is shown in ***Fig***. 8.2. The period of the potential is given by a. The height of the potential energy depends only on one spatial coordinate, namely on x. The potential energy has a value of U_0 for $-c < x < 0$, and a value of zero for $0 \le x \le b$. The potential shown in ***Fig***. 8.2 is periodic with a period a and hence

$$U(x) \quad = \quad U(x+a) \quad = \quad U(x+2a) \quad = \quad \cdots \tag{8.11}$$

In order to obtain the electron states in the one-dimensional potential, the time-independent Schrödinger equation must be solved. Introducing the abbreviations

$$\alpha^2 \quad = \quad 2\,m\,E\,/\,\hbar^2 \tag{8.12}$$

and

$$\beta^2 \quad = \quad 2\,m\,(U_0 - E)\,/\,\hbar^2 \tag{8.13}$$

the time-independent Schrödinger equation can be written inside the well as

$$\frac{d^2\psi}{dx^2} + \alpha^2\,\psi \quad = \quad 0 \qquad\qquad \text{for} \qquad 0 \le x \le b \tag{8.14}$$

and inside the barrier as

$$\frac{\mathrm{d}^2 \psi}{\mathrm{d}x^2} + \beta^2 \psi = 0 \qquad \text{for} \qquad -c < x < 0 \,. \tag{8.15}$$

Kronig and Penney (1930) used the Bloch wave function of Eq. (8.8) for the wave function ψ in the Schrödinger equation. Insertion of the Bloch wave function into the Schrödinger equation and calculating the second derivative with respect to x, *i. e.* $\mathrm{d}^2 / \mathrm{d}x^2$, yields

$$\frac{\mathrm{d}^2 u_{nk}}{\mathrm{d}x^2} + 2\,\mathrm{i}\,k \frac{\mathrm{d}u_{nk}}{\mathrm{d}x} - k^2 u_{nk} + \alpha^2 u_{nk} = 0 \qquad \text{for} \qquad 0 \le x \le b \tag{8.16}$$

and

$$\frac{\mathrm{d}^2 u_{nk}}{\mathrm{d}x^2} + 2\,\mathrm{i}\,k \frac{\mathrm{d}u_{nk}}{\mathrm{d}x} + k^2 u_{nk} + \beta^2 u_{nk} = 0 \qquad \text{for} \qquad -c < x < 0 \tag{8.17}$$

where $k = |\vec{k}\,|$ and $\vec{k}$ was assumed to be pointed along the x direction. The solution of the Schrödinger equation inside the well and inside the barrier are oscillating exponential functions and exponentially decaying functions, respectively. Recall that exponential functions are eigenfunctions of the Schrödinger equation. The solutions are given by

$$u_{nk}(x) = A\,\mathrm{e}^{\mathrm{i}(\alpha - k)x} + B\,\mathrm{e}^{-\mathrm{i}(\alpha + k)x} \qquad \text{for} \qquad 0 \le x \le b \tag{8.18}$$

and

$$u_{nk}(x) = C\,\mathrm{e}^{(\beta - \mathrm{i}k)x} + D\,\mathrm{e}^{-(\beta + \mathrm{i}k)x} \qquad \text{for} \qquad -c < x < 0 \tag{8.19}$$

where A, B, C, and D are four unknown constants. The four constants can be determined by introducing appropriate boundary conditions. At the two boundaries of the potential, *i. e.* at $x = 0$ and $x = b$, the wave function and its derivative must be continuous, that is $u_{nk}(x)$ and $\mathrm{d}u_{nk}(x) / \mathrm{d}x$ must be continuous. These boundary conditions yield the four equations:

Continuity of u_{nk} at $x = 0$:
$$A + B = C + D \tag{8.20}$$

Continuity of u_{nk}' at $x = 0$:
$$\mathrm{i}\,\alpha\,(A - B) = \beta\,(C - D) \tag{8.21}$$

Continuity of u_{nk} at $x = b$:
$$A\,\mathrm{e}^{\mathrm{i}\alpha b} + B\,\mathrm{e}^{-\mathrm{i}\alpha b} = \mathrm{e}^{\mathrm{i}ka}\,\beta\left(C\,\mathrm{e}^{-\beta c} + D\,\mathrm{e}^{\beta c}\right) \tag{8.22}$$

Continuity of u_{nk}' at $x = b$:
$$\mathrm{i}\,\alpha\left(A\,\mathrm{e}^{\mathrm{i}\alpha b} - B\,\mathrm{e}^{-\mathrm{i}\alpha b}\right) = \mathrm{e}^{\mathrm{i}ka}\,\beta\left(C\,\mathrm{e}^{-\beta c} - D\,\mathrm{e}^{\beta c}\right) \tag{8.23}$$

This homogeneous system of linear equations has the four unknowns A, B, C, and D. The system has non-trivial solutions, only if its determinant vanishes. This condition can be expressed as

$$\begin{vmatrix} 1 & 1 & -1 & -1 \\ i\alpha & -i\alpha & -\beta & \beta \\ e^{i\alpha b} & e^{-i\alpha b} & -e^{ika-\beta c} & -e^{ika+\beta c} \\ i\alpha\, e^{i\alpha b} & -i\alpha\, e^{i\alpha b} & -\beta\, e^{ika-\beta c} & \beta\, e^{ika+\beta c} \end{vmatrix} = 0 \qquad (8.24)$$

Evaluation of the determinant yields the condition

$$\frac{\beta^2-\alpha^2}{2\alpha\beta}\sinh(\beta c)\,\sin(\alpha b) + \cosh(\beta c)\,\cos(\alpha b) = \cos(k a). \qquad (8.25)$$

The left side of the equation is a function of E since $\alpha = \alpha(E)$ and $\beta = \beta(E)$ as defined in Eqs. (8.12) and (8.13). Denoting the left side of the equation as $L(E)$, the condition of a vanishing determinant is given by

$$\boxed{L(E) = \cos(k a)} \qquad (8.26)$$

where the function $L(E)$ is given by

$$\begin{aligned} L(E) &= \frac{\beta^2-\alpha^2}{2\alpha\beta}\sinh(\beta c)\,\sin(\alpha b) + \cosh(\beta c)\,\cos(\alpha b) \\ &= \frac{U_0-2E}{2\sqrt{E\,U_0-E^2}}\sinh\left(\sqrt{\frac{2m}{\hbar^2}(U_0-E)}\;c\right)\sin\left(\sqrt{\frac{2m}{\hbar^2}E}\;b\right) \\ &\quad + \cosh\left(\sqrt{\frac{2m}{\hbar^2}(U_0-E)}\;c\right)\cos\left(\sqrt{\frac{2m}{\hbar^2}E}\;b\right) \end{aligned} \qquad (8.27)$$

Equation (8.26) has a solution, only if $|L(E)| \le 1$ because the cosine function on the right-hand side of the equation is limited to values ≤ 1. We will next discuss the function $L(E)$ and differentiate between the regimes in which $L(E) > 1$ and $L(E) \le 1$.

The function $L(E)$ is illustrated in ***Fig***. 8.3. $L(E)$ is an oscillating function whose amplitude decreases with increasing energy. For values of $L(E) > 1$, Eq. (8.26) has no solution. Therefore the Schrödinger equation has no solution for the range of energy which yield $L(E) > 1$. These ranges of energy, for which the Schrödinger equation has no solution, are the *forbidden energies* of the periodic potential considered here. We call the ranges of forbidden energies the ***forbidden energy gap*** of the one-dimensional periodic potential. On the other hand, solutions of the Schrödinger equation are obtained in those ranges of the energy, for which $L(E) \le 1$. We call these ranges of energies the ***allowed energy bands*** of the one-dimensional periodic potential. We thus see that allowed and forbidden ranges of energy, which have been obtained in the previous section from considerations of the chemical bond, also follow from the solution of the Schrödinger equation in a periodic potential.

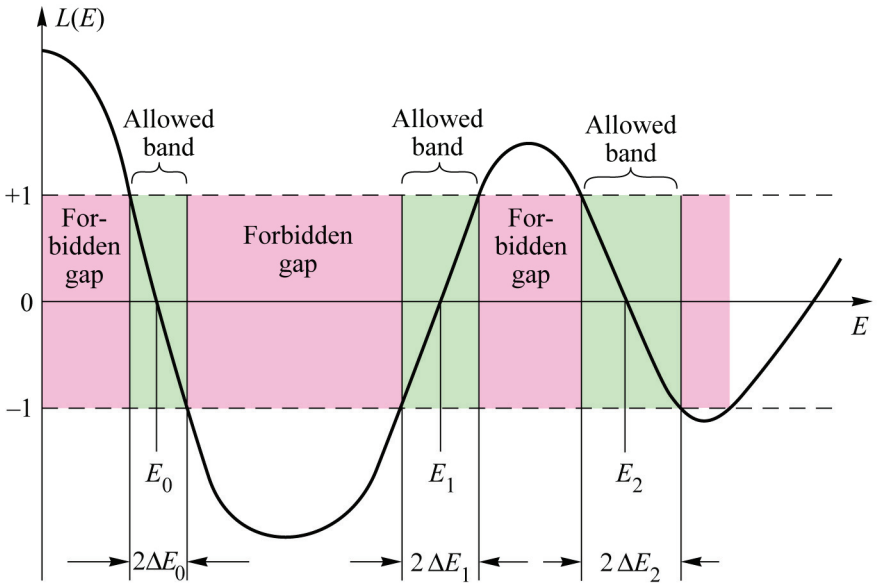


Fig. 8.3 Band structure of a one-dimensional lattice. The function $L(E)$ defines the allowed bands and the forbidden gaps of the lattice. The allowed bands have a center energy of E_n and an energetic width of $2\Delta E_n$. With increasing energy the allowed bands become wider and the forbidden gaps narrower.

As a simple test of Eqs. (8.26) and (8.27), we assume that the thickness of the barrier regions are $c = 0$. Then $a = b$ and the function $L(E)$ is given by

$$L(E) \quad = \quad \cos\left(\sqrt{2\,m\,E/\hbar^2}\;\; b\right). \tag{8.28}$$

Insertion of this result into Eq. (8.26) and using $a = b$ yields

$$E \quad = \quad \frac{\hbar^2\, k^2}{2\,m}\,. \tag{8.29}$$

Thus we have recovered the free particle dispersion relation of Eq. (8.2) in the absence of a periodic potential.

The Kronig–Penney model not only provides the allowed and forbidden bands in a periodic potential, but also the dispersion relation $E(k)$ of an electron propagating in the periodic potential. To derive an analytic form of the dispersion relation, we use the good linearity exhibited by $L(E)$ within the allowed bands. ***Figure*** 8.3 shows a good linearity especially for the allowed band with the lowest energy. If the center of the lowest band is denoted as E_0 and the bandwidth of this lowest band as $2\,\Delta E_0$, then the linearized function $L(E)$ is given by

$$\boxed{L(E) \quad = \quad -\frac{1}{\Delta E_0}\;\left(E - E_0\right)} \tag{8.30}$$

For the next higher energy band, *i. e.* the $n = 1$ band, the function $L(E)$ crosses the band in the opposite direction. The function $L(E)$ is then given by

$$L(E) \quad = \quad \frac{1}{\Delta E_1}\;\left(E - E_1\right). \tag{8.31}$$

For the nth band, the function $L(E)$ is given by

$$L(E) \;=\; (-1)^{n+1}\,\frac{1}{\Delta E_n}\,(E - E_n) \qquad \text{for} \qquad n = 0, 1, 2 \ldots \tag{8.32}$$

Using the linearized forms of $L(E)$ in Eq. (8.26) yields the dispersion relation of an electron in a periodic potential. We obtain for the lowest band

$$\boxed{E \;=\; E_0 \;-\; \Delta E_0 \cos ka} \tag{8.33}$$

For the next higher band, the dispersion relation is given by

$$E \;=\; E_1 \;+\; \Delta E_1 \cos ka\,. \tag{8.34}$$

For the nth band, the dispersion relation is given by

$$E \;=\; E_n \;+\; (-1)^{n+1}\,\Delta E_n \cos ka \qquad \text{for} \qquad n = 0, 1, 2 \ldots \tag{8.35}$$

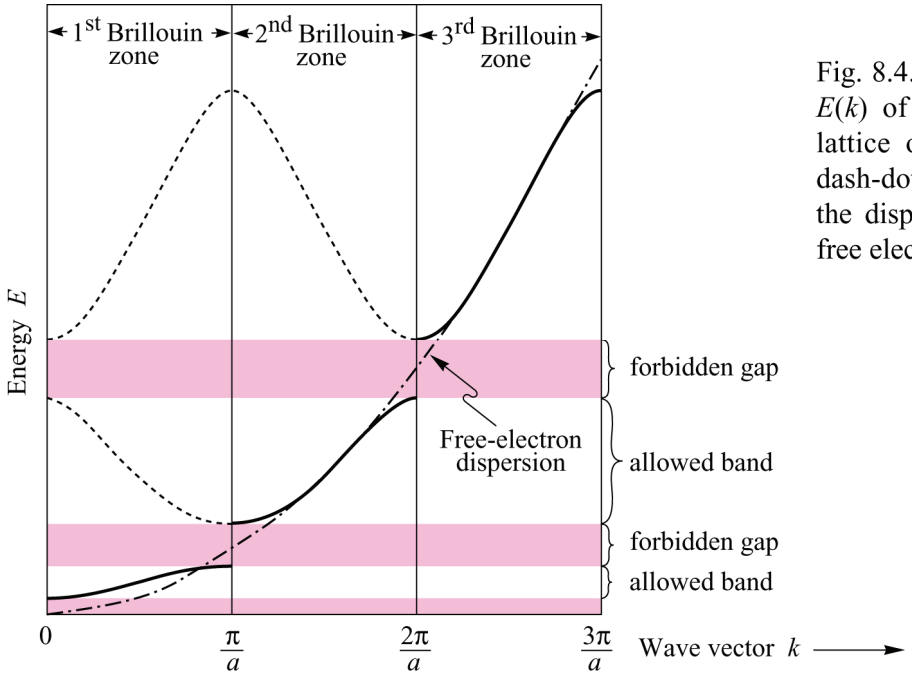


Fig. 8.4. Dispersion relation $E(k)$ of a one-dimensional lattice of a period a. The dash-dotted line represents the dispersion relation of a free electron.

The dispersion relations of the three lowest bands are shown schematically in ***Fig***. 8.4. The dispersion relation is shown only for positive k values. Note that the dispersion relation is an *even* function due to the even characteristic of the cosine function in Eqs. (8.33) to (8.35). Also included is the parabolic dispersion relation (dash-dotted curve) of a free particle. Comparison of the calculated dispersion relation in the one-dimensional potential and the parabolic dispersion relation are very similar around $k = 0$. This is an important result. It shows that an electron in a periodic potential has, near $k = 0$, a similar dispersion relation as a free particle, *i. e*. a parabolic dispersion relation. The comparison also reveals that significant differences between the free particle dispersion and the periodic potential dispersion exist at and near $k = \pm\pi/a, \pm 2\pi/a$ ... that is, when half of the particle wavelength ($\lambda = 2\pi/k$) or integer multiples of one half wavelength is equal to the period of the one-dimensional lattice. This condition is exactly the

Bragg reflection condition. Using $\theta = 90°$ in Eq. (8.6b), the equation reduces to $n\lambda = 2a$ (for $n = 1, 2, 3$ ...) which is identical to the condition stated above. (For a one-dimensional lattice, a normal incidence angle, $\theta = 90°$, must be chosen since particles propagate in normal direction to the potential barriers).

The k interval $-\pi/a \le k \le \pi/a$ is called the ***first Brillouin zone***. The adjacent intervals ranging from $-2\pi/a$ to $-\pi/a$ and from π/a to $2\pi/a$ are the ***second Brillouin zone*** of the one-dimensional periodic potential, and so on.

Figure 8.5 shows the free particle dispersion (dotted line) and the dispersion of a particle in a periodic potential (solid line). Those parts of the dispersion relation that resemble the free particle dispersion relation are shown as solid lines. Inspection of ***Fig***. 8.5 reveals the similarity of the two dispersion relations except for those wave vectors which are integer multiples of π/a.

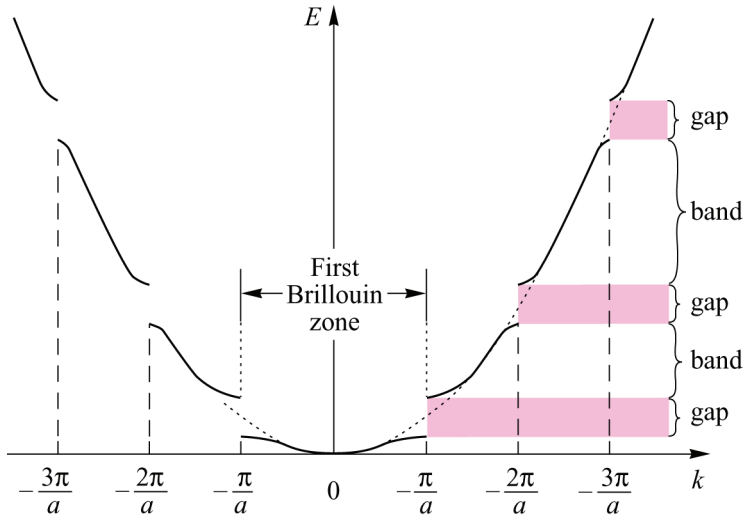


Fig. 8.5. The energy of an electron as a function of its wave vector (*i. e*. dispersion relation) in a one-dimensional periodic potential.

The dispersion relation of a real three-dimensional semiconductor lattice can be much more complex than the simple one-dimensional model considered above. In semiconductors, the dispersion relation depends on the propagation direction of the electron, since the atomic structure and hence the periodic potential depends on the propagation direction of the electron. The dispersion relation for charge carriers in solids is called the energy ***bandstructure*** of the solid. The bandstructure of two important semiconductors, GaAs and Si are shown in ***Fig***. 8.6 (Chelikowski and Cohen, 1976). The center of the Brillouin zone at $k = 0$ is denoted by the Greek letter Γ (Gamma). In GaAs the minimum of the conduction band and the maximum of the valence band occur at the Γ point. Semiconductors in which these two band extrema occur at the Γ point are called ***direct-gap semiconductors***. In direct-gap semiconductors, the electron momentum ($p = \hbar k$) does not change for transitions from the conduction band minimum at $k = 0$ to the valence band maximum at $k = 0$. In Si, the minimum of the conduction band occurs on the Δ axis, whereas the maximum of the valence band is located at the Γ point. Semiconductors in which the minimum of the conduction band occurs at a different k value than the maximum of the valence band, is called an ***indirect-gap semiconductor***. Transitions between the band extrema preserving the momentum of the carrier are impossible in such indirect-gap semiconductor.

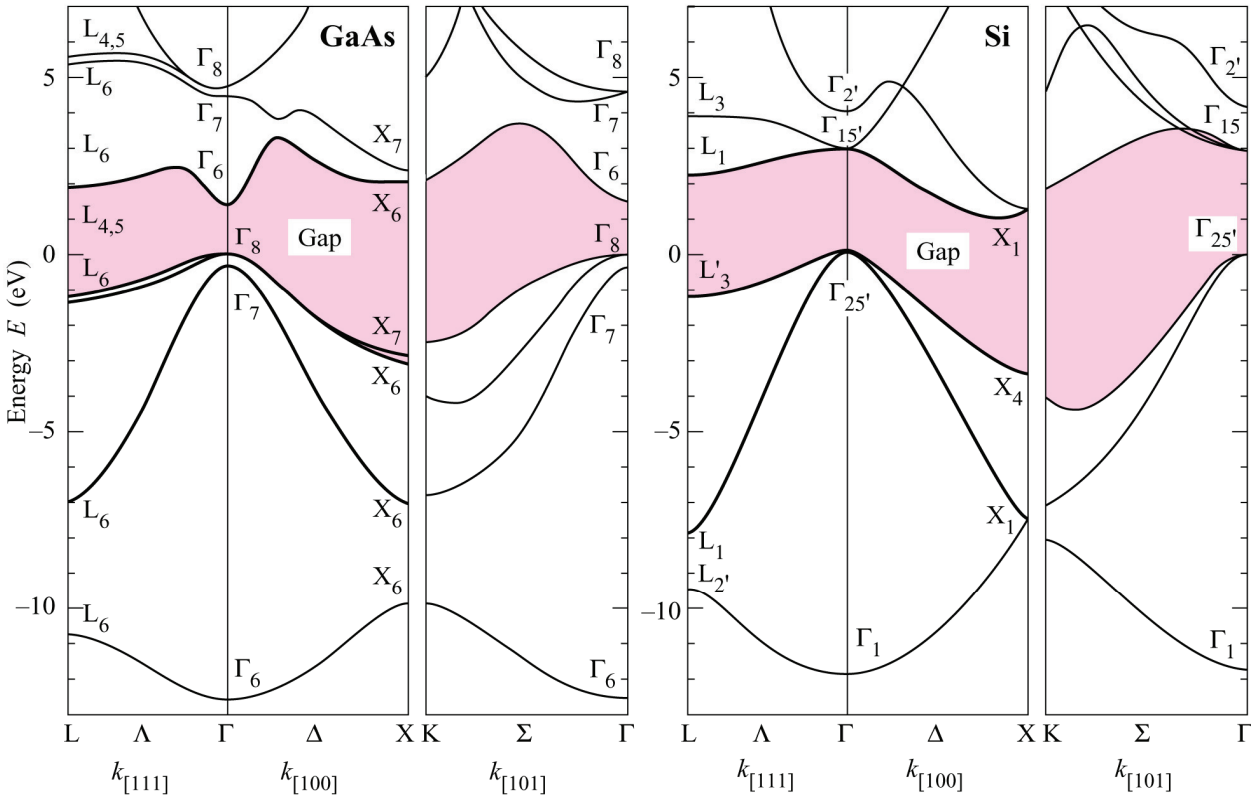
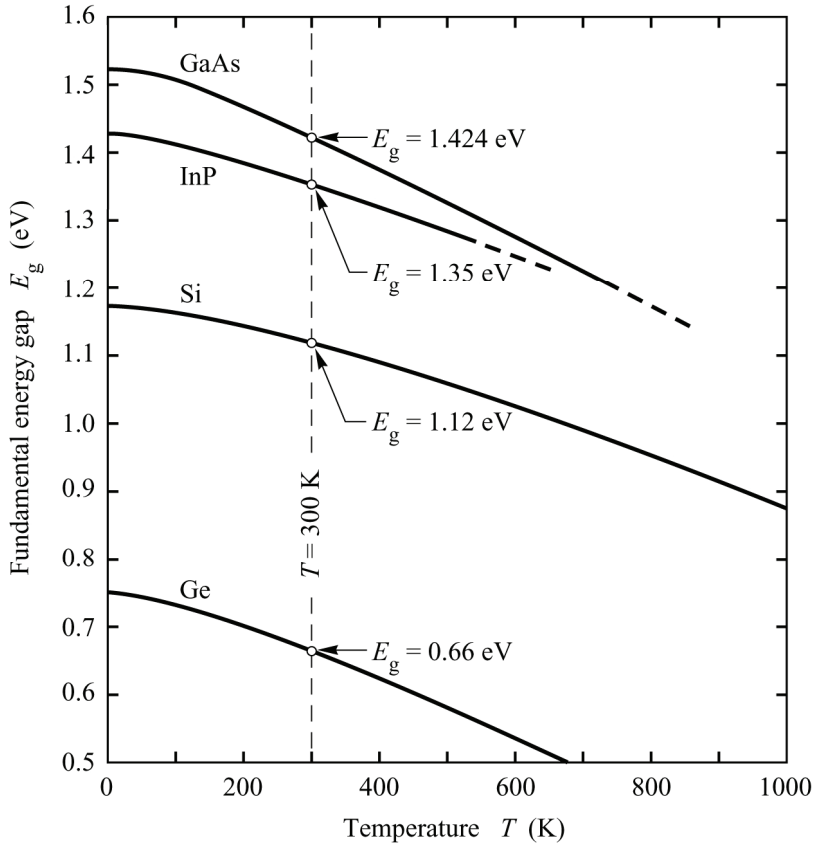


Fig. 8.6 Dispersion relation (band structure) for electrons and holes in the conduction and valence band within the first Brillouin zone for GaAs and Si.

Figure 8.6 reveals that a semiconductor may have several conduction-band minima and valence-band maxima. The energy difference between the lowest conduction band minimum and the highest valence-band maximum is called the fundamental energy gap or ***fundamental gap*** of the semiconductor. The fundamental gaps of GaAs, InP, Si, and Ge are shown in ***Fig***. 8.7 as a function of temperature (Thurmond, 1975; Laufer *et al*., 1980; Pearsall *et al*., 1983). For a discussion of the physics of the temperature dependence of the energy gap, the reader is referred to the literature (Cohen and Chadi, 1980). Here we restrict ourselves to a phenomenological description of the temperature dependence of the energy gap. With good approximation, the fundamental gap can be described by a parabolic dependence on temperature (Varshni, 1967)

$$E_g(T) \quad = \quad E_g(T = 0\,\mathrm{K}) \; - \; \frac{\alpha\, T^2}{T + \beta} \tag{8.36}$$

where $E_g(T = 0\ \mathrm{K})$ is the gap energy at zero temperature and α and β are the two parameters describing the parabolic dependence of the gap energy on temperature. The values of α and β for GaAs, InP, Si, and Ge are given in the inset ***Fig***. 8.7.



$$E_g = E_g(0\text{K}) - \frac{\alpha T^2}{T + \beta}$$

	E_g(0K)	α (10^{-4} eV/K)	β (K)
GaAs	1.519	5.41	204
InP	1.425	4.50	327
Si	1.170	4.73	636
Ge	0.744	4.77	235

Fig. 8.7. Fundamental energy gap of GaAs, InP, Si, and Ge as a function of temperature. The energy gap can be approximated by a parabolic equation with the fitting parameters α and β.

8.4 The effective mass

The influence of the periodic potential experienced by an electron propagating in an atomic lattice can be taken into account by the elegant and powerful concept of the effective mass. As will be seen, the mass of a charge carrier, *e. g*. the mass of an electron of $m_0 = 9.11 \times 10^{-31}$ kg, is modified to an effective value m^* due to the influence of the periodic potential. By this modification, the entire influence of the periodic potential is taken into account. That is, electrons in the periodic potential with the effective mass m^* can be treated as free electrons.

To derive the effective mass of a charge carrier in a periodic potential, we use the definition of the effective mass in Newton's second law and then apply this definition to a quantum-mechanical wave-like particle whose dispersion relation is assumed to be known. Newton's second law defines the mass of a particle in terms of the acceleration a of the particle caused by a force F acting on the particle.

$$F = m a \tag{8.37}$$

The acceleration can be expressed as a change of the group velocity of the quantum-mechanical wave representing the particle, that is $a = (\mathrm{d}/\mathrm{d}t)\, v_{gr} = (\mathrm{d}/\mathrm{d}t)\, \mathrm{d}\omega/\mathrm{d}k$, where ω is the angular frequency of oscillation of the wave. Assuming that the particle has only kinetic energy, the energy of the particle is given by Planck's relation $E = \hbar\,\omega$. Hence the mass in Eq. (8.37) can be expressed as

$$a = \frac{1}{\hbar}\,\frac{\mathrm{d}^2 E}{\mathrm{d}t\,\mathrm{d}k} = \frac{1}{\hbar}\,\frac{\mathrm{d}^2 E}{\mathrm{d}k^2}\,\frac{\mathrm{d}k}{\mathrm{d}t}\,. \tag{8.38}$$

Using the de Broglie relation ($p = \hbar\, k$), Newton's second law can be expressed as

$$F = \frac{\mathrm{d}p}{\mathrm{d}t} = \hbar \frac{\mathrm{d}k}{\mathrm{d}t}. \tag{8.39}$$

This relation is true for a free electron. It is also valid for electrons in any potential including the periodic potential. Insertion of Eq. (8.39) into Eq. (8.38) yields the acceleration as a function of the E-versus-k relation

$$a = \frac{1}{\hbar^2} \frac{\mathrm{d}^2 E}{\mathrm{d}k^2} F. \tag{8.40}$$

Comparison of this equation with Newton's second law ($a = F/m$) allows us to express the mass of an electron in a periodic potential as

$$\boxed{m^* = \frac{\hbar^2}{\mathrm{d}^2 E / \mathrm{d}k^2}} \tag{8.41}$$

The mass given by Eq. (8.41) is called the ***effective mass*** of a charge carrier. Depending on the nature of the periodic potential, the effective mass may be lighter or heavier than the free electron mass.

According to Eq. (8.41), the ***effective mass is inversely proportional to the second derivative of E with respect to k***, that is, the ***effective mass is inversely proportional to the curvature of the dispersion relation***. A strongly curved $E(k)$ relation implies a small effective mass, whereas a weakly curved dispersion relation indicates a heavy effective mass. In the case of a parabolic dispersion relation, the second derivative of E with respect to k is a constant. As a consequence, the effective mass is a constant as well, that is, the effective mass has a constant value independent of energy.

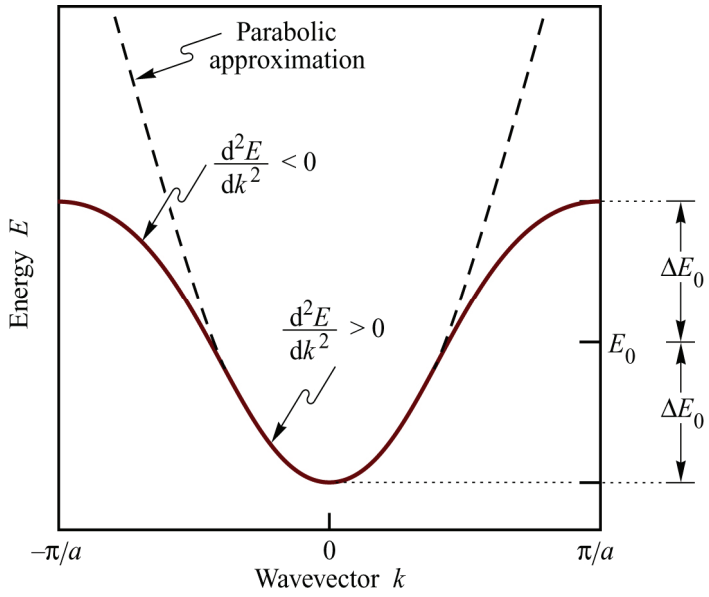


Fig. 8.8. Dispersion relation of a one-dimensional lattice with positive effective mass ($\mathrm{d}^2E/\mathrm{d}k^2 > 0$) near the zone-center, and negative effective mass ($\mathrm{d}^2E/\mathrm{d}k^2 < 0$) near the zone boundary. Close to the zone center, the dispersion relation can be approximated by a parabola.

Generally, the effective mass depends on energy. ***Figure*** 8.8 illustrates the basic shape of the dispersion relation within the first Brillouin zone. The curvature of the $E(k)$ relation depends on k and therefore also on E. In the vicinity of the zone center ($k \approx 0$), the second derivative of $E(k)$

is positive and consequently the effective mass is also positive. For increasing values of k, the curvature of $E(k)$ deviates from the parabolic dispersion (dashed line in ***Fig***. 8.8) and the curvature becomes negative close to the edge of the first Brillouin zone. Hence, the effective mass assumes negative values as well.

What is the physical meaning of a negative effective mass? Let us consider an electron with $k = 0$, which is accelerated by a constant force F. Close to the Brillouin-zone center, the electron behaves identical to a free electron with the effective mass m^*. As the electron assumes increasingly higher values of k, the interaction with the lattice becomes stronger. At a k value of $k = \pi/(2a)$, that is half way between zone center and zone edge, the curvature of $E(k)$ is zero, *i. e*. the effective mass is infinity heavy. This means that the velocity of the electron cannot be further increased by the force F. That is, the change in group velocity ($\mathrm{d}^2\omega/\mathrm{d}k^2$) equals zero at the half-way point. As the zone boundary is approached, the mass becomes negative, *i. e*. the electron is accelerated in the opposite direction of the force F. When the zone boundary is reached, the group velocity of the particle equals zero, since $\mathrm{d}\omega/\mathrm{d}k = (1/\hbar)\,\mathrm{d}E/\mathrm{d}k = 0$. Even though the group velocity of the electron is zero, the momentum of the electron is finite and it is given by $p = \hbar k$. The electron can be thought to be represented by two waves, one propagating in positive k direction and a second identical wave propagating in the negative k direction. Thus the electron is represented by a standing wave with a zero group velocity.

The periodic potential of a crystal depends on the propagation direction. As a consequence, the dispersion relation and the effective mass also depend on the direction of propagation. Generally, the effective mass is a tensor and not just a scalar. Newton's second law is then given by

$$\begin{pmatrix} F_\mathrm{x} \\ F_\mathrm{y} \\ F_\mathrm{z} \end{pmatrix} = \begin{pmatrix} m^*_\mathrm{xx} & m^*_\mathrm{xy} & m^*_\mathrm{xz} \\ m^*_\mathrm{yx} & m^*_\mathrm{yy} & m^*_\mathrm{yz} \\ m^*_\mathrm{zx} & m^*_\mathrm{zy} & m^*_\mathrm{zz} \end{pmatrix} \begin{pmatrix} a_\mathrm{x} \\ a_\mathrm{y} \\ a_\mathrm{z} \end{pmatrix} \tag{8.42}$$

Consider a force along the x direction acting on an electron. Depending on the band structure of the semiconductor, the electron may be accelerated along a direction normal to the x direction, *e. g*. the y direction. In analogy to Eq. (8.38), the acceleration along the y direction is given by

$$a_y = \frac{1}{\hbar}\,\frac{\partial^2 E}{\partial k_\mathrm{y}\,\partial t} = \frac{1}{\hbar}\,\frac{\partial^2 E}{\partial k_\mathrm{y}\,\partial k_\mathrm{x}}\,\frac{\partial k_\mathrm{x}}{\partial t}\,. \tag{8.43}$$

We assumed that the force is directed along the x direction and Eq. (8.39) is then given by

$$F_\mathrm{x} = \frac{\mathrm{d}p_\mathrm{x}}{\mathrm{d}t} = \hbar\,\frac{\mathrm{d}k_\mathrm{x}}{\mathrm{d}t}\,. \tag{8.44}$$

Insertion of Eq. (8.44) into Eq. (8.43) allows us to identify the tensor element m_xy^* as

$$\boxed{m^*_\mathrm{xy} = \hbar^2 \left(\frac{\partial^2 E}{\partial k_\mathrm{x}\,\partial k_\mathrm{y}}\right)^{-1}} \tag{8.45}$$

The first subscript x of the mass m_{xy}* refers to the direction of the force, whereas the second subscript refers to the direction of the acceleration.

For semiconductors with an isotropic dispersion relation with a band minimum at $k = 0$, the effective mass tensor has only diagonal tensor elements and no off-diagonal elements, *i. e.* m_{ij}* = 0 for $i \neq j$. In the case of an isotropic semiconductor it is m_{xx}* = m_{yy}* = m_{zz}* and thus the effective mass tensor reduces to a scalar. As an example, we consider GaAs which has an isotropic band structure with a conduction band minimum at $k = 0$. As a consequence, the effective mass is a scalar, *i. e.* independent of the propagation direction.

We next consider the effective mass in the Kronig–Penney model. The dispersion relation in the Kronig–Penney model is given by Eqs. (8.33) to (8.35). The dispersion of the lowest band is given by

$$E = E_0 - \Delta E_0 \cos ka \,. \tag{8.46}$$

Expanding the cosine function into a power series yields

$$E = E_0 - \Delta E_0 \left(1 - \frac{(k\,a)^2}{2!} + \frac{(k\,a)^4}{4!} - \cdots \right). \tag{8.47}$$

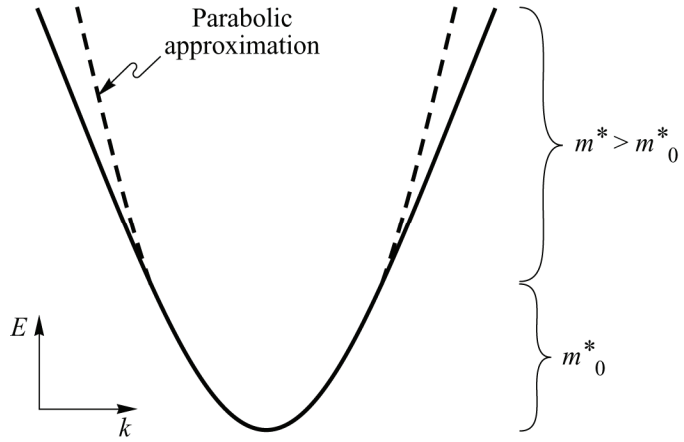


Fig. 8.9. Parabolic dispersion (dashed curve) and deviations from parabolic dispersion at high energies. The non-parabolic dispersion leads to an increase of the effective mass at high energies.

The dispersion relation is schematically shown in ***Fig***. 8.9. Near $k = 0$, the term $(k\,a)^2$ dominates and the dispersion relation is parabolic. For larger values of k, the term $(k\,a)^4$ cannot be neglected and the dispersion deviates from the parabolic dependence. Consider now the dispersion relation near the energy minimum, *i. e.* near the bottom of the band. In the vicinity of $k = 0$, the term $(k\,a)^4 / 4!$ and all higher terms can be neglected. Using Eq. (8.41), the effective mass near the bottom of the band is given by

$$\boxed{m_0^* = \frac{\hbar^2}{\Delta E_0\, a^2}} \tag{8.48}$$

where the subscript "0" in m_0^* indicates that this effective mass is valid in the vicinity of the zone center near $k = 0$. The effective mass at the bottom of a miniband is denoted as the

confinement mass. Equation (8.48) further shows that the effective mass is heavy for small bandwidths ΔE_0. For thick and / or high barriers in the periodic potential, the bandwidth becomes very small. In such potentials, the tunneling probability through the barrier becomes small as well. That is, the transfer of electrons from one well to the next well is strongly impeded by the thick and / or high barriers. This situation can be understood in terms of a large effective mass. Electrons with a large effective mass cannot propagate easily in a periodic potential.

The minimum of the band whose dispersion relation is given by Eq. (8.47) occurs at an energy $E = E_0 - \Delta E_0$. At the bottom of the band the effective mass is given by Eq. (8.48). However, at higher energies, the dispersion relation is no longer parabolic and, as a consequence, the effective mass changes. The energy dependence of the effective mass plays an important role in several areas of semiconductor heterostructures. For example, the quantization of energy levels in quantum wells depends on the effective mass of the carriers. To calculate the change of the effective mass due to the non-parabolicity of the band structure, we employ the terms $(k\,a)^2$ and $(k\,a)^4$ in the dispersion relation of Eq. (8.47) and neglect all higher-order terms. The term $(k\,a)^4$ in Eq. (8.47) has the opposite sign of the term $(k\,a)^2$ and therefore the former term reduces the curvature of the dispersion relation. Consequently, the effective mass will *increase* for higher energies. Calculation of the effective mass from the dispersion relation by using Eq. (8.45) yields

$$m^* \;=\; m_0^* \left(1 + \frac{E - \left(E_0 - \Delta E_0\right)}{\Delta E_0} \right) \tag{8.49}$$

where m_0* is the effective mass at the bottom of the band as given in Eq. (8.48). The bottom of the band occurs at the energy $E_0 - \Delta E_0$. The bandwidth of the band is $2\,\Delta E_0$. Hence Eq. (8.49) indicates that the effective mass increases over m_0* for higher energies. Equation (8.48) shows that the approximation $m^* = m_0^*$ is valid near the bottom of the band, specifically for energies $[E - (E_0 - \Delta E_0)] << \Delta E_0$.

Exercise 1: ***The Kronig–Penney model***, ***the dispersion relation, and the effective mass***. A periodic potential consists of 1.0 nm wells and 2.0 nm barriers with a barrier height of 200 meV. Using the Kronig–Penney model, calculate the number of bands and the respective dispersion relations, assuming that the carrier mass in the absence of the periodic potential is the free electron mass m_e. Using appropriate approximations, calculate the band widths and the effective masses in the bands. Explain the trend found for ΔE_n and m_e* as the band index n increases.

Solution: Graphical solution of the eigenvalue equation $L(E) = \cos ka$ reveals that there is only one band between 0 and 200 meV with E_0 = 97.4 meV and ΔE_0 = 4.2 meV. The dispersion relation is thus given by E = 97.4 meV + 4.2 meV cos (30 Å k). The width of this band is $2\Delta E_0$ = 8.4 meV. The effective mass in the lowest band is $m_e^* = 1.84 \times 10^{-30}$ kg = 2.02 m_e. The following trends are generally found as the band index n increases: Band width ΔE_n increases; effective mass $m_{e,\,n}$* decreases.

How do properties of the *bandstructure*, *i. e*. the number of bands, their widths, and the effective masses change, as the barrier width decreases? How does the bandstructure of the lowest band change as the barrier width increases? What is the value of the lowest band width and effective mass in the limit of infinitely thick barriers?

Solution: As barrier width decreases, the number of bands increases, band widths increase, and effective masses decrease. As the barrier width increases, the band width of the lowest band decreases and the dispersion relation becomes less curved. In the limit of infinitely thick barriers, one obtains $\Delta E_0 \to 0$ and $m_{e,0}^* \to \infty$.

Exercise 2: ***The effective mass in superlattices***. *Semiconductor superlattices* are periodic structures consisting of two different semiconductors. Semiconductors already have bands (*e. g.* the conduction band) and thus we denote the newly formed bands as "*minibands*". Consider an $Al_xGa_{1-x}As/GaAs$ superlattice with a period of 4.0 nm and a barrier width of 2.0 nm and height of 250 meV. Calculate the properties of this superlattice including the number of minibands, energies, band widths, and effective confinement masses. Assume that the effective mass in bulk GaAs and $Al_xGa_{1-x}As$ is $m_e^* = 0.067 \times m_e$.

Next consider that the electrons move in parallel direction to the layers of the superlattice. What effective mass do you expect for motion parallel to the superlattice layers? Is transport in these superlattice structures isotropic or anisotropic?

Solution: There is one miniband between 0 and 250 meV (there is an additional allowed band between 190 meV and 250 meV, but it is only partially within the well). From the calculation we obtain $E_0 = 23.7$ meV and $\Delta E_0 = 0.4$ meV. The bandwidth is $2\Delta E_0 = 0.8$ meV. Therefore, the dispersion relation for the lowest miniband is given by $E = 23.7\,\text{meV} - 0.4\,\text{meV} \cos(4.0\,\text{nm}\, k)$.

The effective mass at the bottom of the miniband is given by

$$m^* = \frac{\hbar^2}{\mathrm{d}^2 E / \mathrm{d}k^2} = \frac{\hbar^2}{\Delta E_0\, a^2} = 108.7 \times 10^{-31}\,\text{kg} = 11.9 \times m_e = 178.3 \times m_e^* .$$

When the electron propagates parallel to the superlattice layers, the effective mass is equal to m_e^* ($= 0.067 \times m_e$) because the periodic potential felt by the electron is the one of the GaAs lattice. Transport in superlattice structures is anisotropic.

Exercise 3: ***Comparison of photon and electron momenta***. Assume that radiative recombination processes occur in GaAs. Calculate wavelength and momentum of a photon with energy 1.42 eV. In optical transitions, momenta must be conserved. Compare the photon momentum with the momentum of an electron located at the boundary of the first Brillouin zone. What conclusions can be drawn from this comparison?

Solution:

Photon wavelength	$\lambda = h\,c / E$	$= 870$ nm
Photon momentum	$p = h / \lambda$	$= 8.0 \times 10^{-28}$ kg m / s
Electron momentum at zone boundary	$p_{max} = \hbar\,\pi / a_0$	$= 5.9 \times 10^{-25}$ kg m / s

The comparison reveals that the momentum difference between electrons and holes taking part in a radiative recombination event must be much smaller than the maximum momentum at the zone boundary. Therefore, optical transitions are "vertical" in *k*-space. Carriers taking part in radiative recombination are usually located near the center of the first Brillouin zone,

i. e. near $k = 0$. The comparison also shows that radiative recombination cannot occur in indirect-gap semiconductors, which have a substantial difference between the location of conduction band minimum and valence band maximum in k space.

8.5 The Bloch oscillation

In the preceding Section, we have learned that an *energy gap* occurs at the electron wave number assumes values of $k = \pm \pi / a, \pm 2\pi / a$ ... . We will next discuss the physical interpretation of these energy gaps. To do this we consider the electron k value of π / a. At this value of k, the electron de Broglie wavelength is given by $\lambda = 2a$.

When an electron propagates in the periodic Kronig–Penney potential, it must tunnel through the barriers. The tunneling probability through a barrier is always less than one, *i. e.* $T < 1$ where the tunneling probability through one barrier is denoted as T. Thus the electron wave gets partially reflected upon propagation through one barrier.

If the electron wave propagating in the Kronig–Penney potential gets indeed partially reflected at each barrier, then the interference of the partially reflected waves is of importance. The difference in path length for waves partially reflected from two adjacent barriers (or *crystal planes*) is $2a$. Since the electron wavelength is $\lambda = 2a$ as well, all partially reflected waves interfere constructively with themselves. This makes the forward propagation of waves with $k = \pi / a$ impossible. As a consequence, the electron wave undergoes what is called ***Bragg reflection***. Bragg reflection is an elastic process in which the *magnitude* of k is *conserved* but the *direction* of k is *reversed*. Similar considerations hold true for electrons with $k = \pm \pi / a$, $\pm 2\pi / a$ ... .

The Bragg reflection can also be understood by taking into account the crystal wave number G. When the electron reaches the zone boundary, it undergoes Bragg reflection and the k' vector of the reflected electron is given by the Bragg reflection condition

$$k' = k + G = k - \frac{2\pi}{a} = -\frac{\pi}{a} \tag{8.50}$$

Hence, the electron is Bragg reflected to the negative side of the zone boundary located at $k = -\pi / a$, without change in total energy.

Now consider an electron that is subjected to an electric field. The electric field exerts a force $F = -e\mathcal{E}$ on the electron. Assume that the electron is initially not in motion, *i. e.* $k = 0$. Upon application of the electric field, the k value of the electron increases from $k = 0$ to π / a. At this value of k, Bragg reflection occurs, and the electron assumes a k value of $-\pi / a$. Then the electron is again accelerated to $k = \pi / a$. At this point, the electron again undergoes Bragg reflection and the cycle starts from the beginning. The process described above is called the ***Bloch oscillation*** of the electron in an energy band of a solid state crystal.

Next we consider the propagation of the electron along the k axis. Recall that the group velocity is given by

$$v_{gr} = \frac{d\omega}{dk} = \frac{1}{\hbar}\frac{dE}{dk} \tag{8.51}$$

The rate of kinetic energy gain of an electron propagating along the x axis in an electric field is given by

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\mathrm{d}(F\,x)}{\mathrm{d}t} = -e\,\mathcal{E}\,\frac{\mathrm{d}x}{\mathrm{d}t} = -e\,\mathcal{E}\,v_{\mathrm{gr}}\ . \tag{8.52}$$

With Eq (8.51) we can write

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\mathrm{d}E}{\mathrm{d}k}\frac{\mathrm{d}k}{\mathrm{d}t} = \hbar\,v_{\mathrm{gr}}\,\frac{\mathrm{d}k}{\mathrm{d}t}\ . \tag{8.53}$$

Equating Eqs. (8.52) and (8.53) yields the rate of change of the k value of the electron according to

$$\boxed{\frac{\mathrm{d}k}{\mathrm{d}t} = -\frac{1}{\hbar}\,e\,\mathcal{E}}\ . \tag{8.54}$$

which is called the ***acceleration theorem*** of electrons in a periodic potential accelerated by an electric field. Note that in a *constant electric field*, *the rate of change for k*, *i. e.* $\mathrm{d}k/\mathrm{d}t$, *is a constant*. Thus, the electron "moves" along the k axis at a *constant* rate.

Exercise 4: ***The period of Bloch oscillations***. Show that the period of the Bloch oscillation is given by

$$\boxed{T_{\mathrm{Bloch}} = \frac{2\,\pi\,\hbar}{e\,\left|\mathcal{E}\right|\,a}} \tag{8.55}$$

Solution: Using $\mathrm{d}k/\mathrm{d}t = (k_{\max} - k_{\min})/T_{\mathrm{Bloch}} = [\pi/a - (-\pi/a)]/T_{\mathrm{Bloch}} = (2\pi/a)/T_{\mathrm{Bloch}}$, and the acceleration theorem, one obtains the result given in Eq. (8.55)

Scattering mechanisms other than Bragg scattering have not been considered in the above discussion. Any inelastic scattering mechanism, *e. g.* phonon scattering, reduces the electron momentum and one can assume that $k \approx 0$. Thus it is difficult for the electron to complete the entire cycle of the Bloch oscillation. Typical inelastic scattering times are 10^{-11} s = 10 ps at low fields and 10^{-13} s = 0.1 ps at high fields.

Exercise 5: ***Bloch oscillation***. Calculate the period of the Bloch oscillation for a = 0.5 nm and $\mathcal{E}$ = 1000 V / cm. Compare the period with typical inelastic scattering times. What conclusions do you draw from the comparison?

Solution: The Bloch oscillation period is calculated to be $T_{\mathrm{Bloch}} = 8.2 \times 10^{-11}$ s. Since T_{Bloch} is much larger than the typical inelastic scattering time, it is unlikely that the electron undergoes a complete Bloch oscillation. Inelastic scattering events are much more likely than elastic Bragg scattering events.

If the electron could reach the zone boundary, it could also overcome the energy gap to the next higher band and then propagate at larger k values in the next higher band. However, such a

transition would require additional energy which would have to be provided. Therefore, Bragg reflection is the more likely process if the electron ever reaches the zone boundary. Bloch oscillations are a theoretically postulated concept but such Bloch oscillations have not been observed experimentally.

Exercise 6: ***Bloch oscillation***. ***Figure*** 8.10 shows the dispersion relation, the electron momentum, and the group velocity along the x direction as a function of time, when an electric field along the negative x direction) is applied to a crystal containing free electrons. Explain ***Figs***. 8.10 (b) and (c). The dashed line in ***Fig***. 8.10 (c) indicates a *linear* dependence of the group velocity near $v_{gr} = 0$. Explain why v_{gr} depends *linearly* on t near $k = 0$.

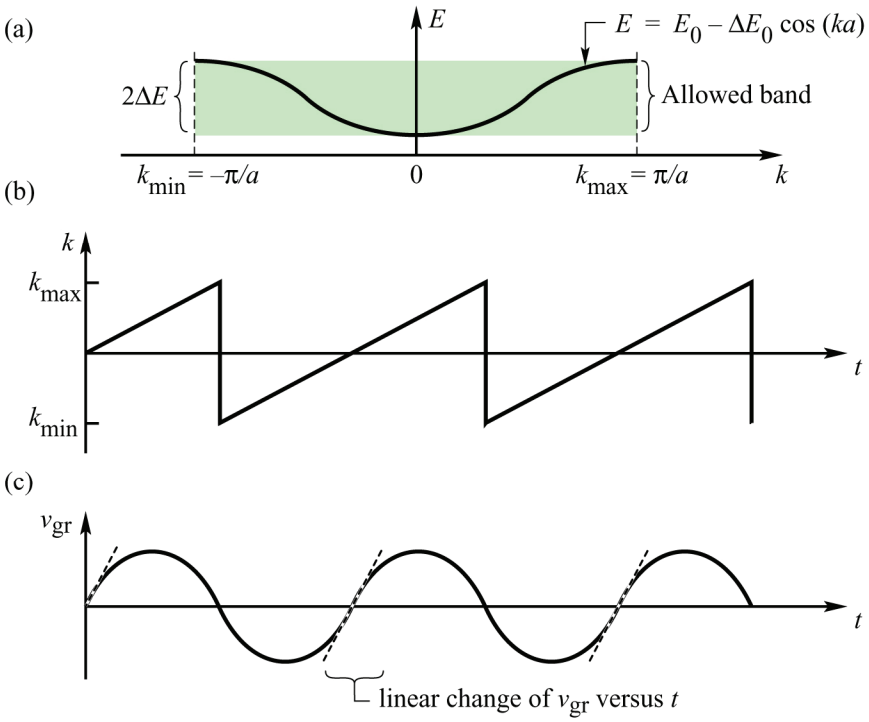


Fig. 8.10. (a) Dispersion relation for electrons in first Brillouin zone of a one-dimensional periodic potential. (b) k versus time and (c) group velocity versus time during Bloch oscillation.

Solution: According to Eq. (8.54), we have $\mathrm{d}k/\mathrm{d}t = -\hbar^{-1} e \mathcal{E}$. Since the electric field $\mathcal{E}$ is constant, and $\mathcal{E} < 0$ (along negative direction) it is $k = -\hbar^{-1} e \mathcal{E} t = C t$ with $C > 0$. Thus, electrons move along the k axis at a constant rate. When an electron reaches the boundary, it is Bragg reflected to the negative side of the boundary, as shown in ***Fig***. 8.10(b). Furthermore the electron energy E changes according to

$$\frac{\mathrm{d}E}{\mathrm{d}k} = \frac{\mathrm{d}(\hbar\omega)}{\mathrm{d}k} = \hbar\frac{\mathrm{d}\omega}{\mathrm{d}k} = \hbar v_{gr} = \frac{\mathrm{d}}{\mathrm{d}k}\left(E_0 - \Delta E_0 \cos(ka)\right) = a\,\Delta E_0 \sin(ka)$$

Therefore,

$$v_{gr} = \frac{a\,\Delta E_0 \sin(ka)}{\hbar} = \frac{a\,\Delta E_0 \sin(Cat)}{\hbar}$$

This is what we see in ***Fig***. 8.10(c), the group velocity of an electron varies sinusoidally with time. At the points where v_{gr} is near zero, the sin function can be approximated by $(\sin x) \approx x$ and one obtains

$$v_{gr} \approx \frac{a\,\Delta E_0\,Ca}{\hbar}\,t$$

That is v_{gr} changes *linearly* with time implicating a constant acceleration. Such constant acceleration in a constant electric field elucidates the similarity between a free electron and an electron in a periodic potential.

8.6 Semiconductor superlattices

As shown in the Kronig–Penney model, bands of allowed states and bands of disallowed states form for electrons propagating in a periodic potential. Semiconductor superlattices are periodic semiconductor structures consisting of two semiconductors with different bandgap. Usually, the periods of semiconductor superlattices are longer as the lattice constant of the constituent semiconductors. As a result, the allowed bands have a narrower width and these bands are therefore called ***minibands***. A miniband in the conduction band of a semiconductor superlattice is schematically shown in ***Fig***. 8.11 (a).

In the absence of an electric field, the energy states of carriers in a superlattice can be calculated by the Kronig–Penney model, taking into account the modified boundary conditions for semiconductor structures as discussed in the preceding section.

However, if an electric field is applied to the semiconductor structure, transport can either proceed via ***miniband conduction*** or by ***sequential tunneling*** depending on the magnitude of the electric field.

Consider the case in which the energy drop due to an electric field per period of the superlattice is less than the miniband width, *i. e.*

$$\left| e\,\mathcal{E}\,a \right| \;<\; 2\,\Delta E \tag{8.56}$$

where $|\, e\,\mathcal{E}\,a\,|$ is the energy drop occurring within one period of the superlattice and $2\Delta E$ is the width of the miniband. In this case, electrons will propagate within the miniband formed by the superlattice. This situation is schematically shown in ***Fig***. 8.11(b)

Next consider the case in which the energy drop due to the electric field per period of the superlattice is larger than the miniband width, *i. e.*

$$\left| e\,\mathcal{E}\,a \right| \;>\; 2\,\Delta E\ . \tag{8.57}$$

In this case, the miniband no longer exists since the structure has lost its strict periodicity. *Discrete levels* rather than a *miniband* will form in each quantum well. Electrons propagate in the superlattice by ***sequentially tunneling*** through the barriers rather than by miniband conduction. This situation is shown in ***Fig***. 8.11(c). The sequence of energy levels is frequently referred to as a ***Stark ladder***, reminiscent of the Stark effect which describes the change of energy levels under the influence of electric fields.

The transport parallel to the superlattice layers is not affected by the superlattice structures since the periodic potential felt by an electron is either the potential of the well material or that of the barrier material, depending on the layer of propagation. Most carriers will propagate in the well layers due to the lower energy of the well layers.

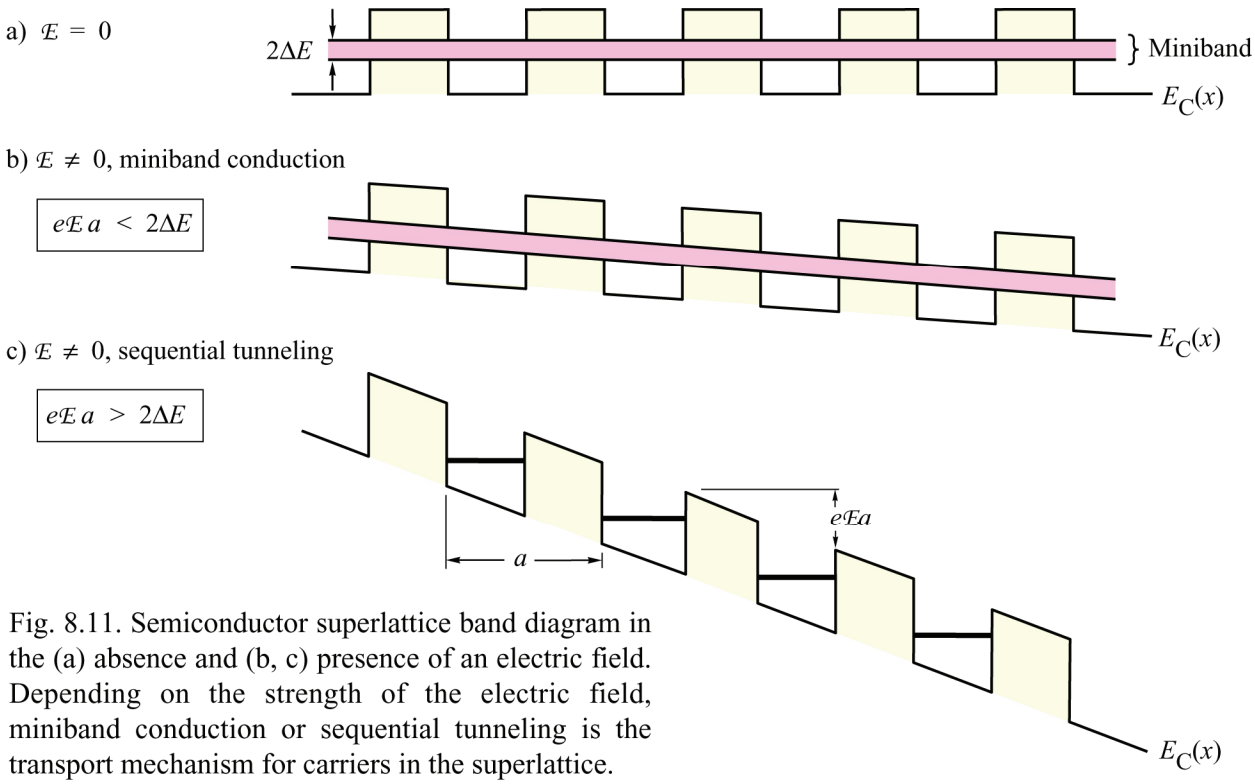


Fig. 8.11. Semiconductor superlattice band diagram in the (a) absence and (b, c) presence of an electric field. Depending on the strength of the electric field, miniband conduction or sequential tunneling is the transport mechanism for carriers in the superlattice.

References

Ashcroft R. L. and Mermin N. D. *Solid State Physics* (Saunders College, Philadelphia, 1976)

Bloch F. "Uber die Quantenmechanik der Elektronen in Kristallgittern" (translated title: "About the quantum mechanics of electrons in crystal lattices") *Zeitschrift für Physik* **52**, 555 (1928)

Bloch F. "Zum elektrischen widerstandsgesetz bei tiefen temperaturen" (translated title: "About electrical resistance law at low temperatures"), *Zeitschrift für Physik* **59**, 208 *(*1930)

Chelikowski J. R. and Cohen M. L. "Nonlocal pseudo-potential calculations for the electronic structure of eleven diamond and zincblende semiconductors" *Physical Review B***14**, 556 (1976)

Cohen M. L. and Chadi D. J. in *Handbook on Semiconductors*, Vol. **2**, P.155, edited by T. S. Moss (North-Holland, Amsterdam, 1980)

Kronig R. D. and Penney W. G. "Quantum mechanics of electrons in crystal lattices" *Proceedings of the Royal Society, London* **A130**, 499 (1930)

Laufer P. M., Pollak F. H., Nahory R. E. and Pollak M. A. "Electro-reflectance investigation of $In_{1-x}Ga_xAs_yP_{1-y}$ lattice-matched to InP" *Solid State Communications* **36**, 419 (1980)

Pearsall T. P., Eaves L., and Portal J. C. "Photoluminescence and impurity concentration in $Ga_xIn_{1-x}As_yP_{1-y}$ alloys lattice-matched to InP" *Journal Applied Physics* **54**, 1037 (1983)

Thurmond C. D. "Standard thermodynamic functions for the formation of electrons and holes in Ge, Si, GaAs, and GaP" *Journal of the Electrochemical Society* **122**, 1133 (1975)

Varshni Y. P "Temperature dependence of the energy gap in semiconductors" *Physica* **34**, 149 (1967)

Felix Bloch (1905–1983)
Developed wave function of electron in a periodic potential

9

Approximate solutions of the Schrödinger equation

9.1 The WKB method

The Schrödinger equation has analytic solutions only for very few selected potential energies $U(x)$. For example, the infinite square well has an analytic solution. If the one-dimensional potential energy does not have a very simple form, the solution of the one-dimensional time-independent Schrödinger equation

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) + U(x)\,\psi(x) = E\,\psi(x) \tag{9.1}$$

is generally a complicated problem. Some approximate methods to solve the Schrödinger equation are the perturbation method (see Chaps. 10 and 11) or the variational method (see Chap. 9). An approximate method of great versatility has been developed by ***Wentzel, Kramers and Brillouin*** (1926) and is called the WKB method or WKB approximation. This method provides approximate wave functions in one-dimensional problems. The WKB method can also be applied to three-dimensional problems, where the potential is spherically symmetric and a radial differential equation can be separated from the three-dimensional problem (see, for example, Bohm, 1951).

The WKB approximation can be used, if the potential energy $U(x)$ varies *slowly*. Specifically, changes in $U(x)$ should be small on the length scale of the de Broglie wavelength. In a constant potential, the Schrödinger equation has the solutions $\exp(\pm \mathrm{i}\,k\,x)$, with $k = 2\pi/\lambda = \text{const}$. If $U(x)$ varies slowly with x, we write the solution in the form

$$\boxed{\psi(x) = \mathrm{e}^{\mathrm{i}\phi(x)}} \tag{9.2}$$

where the function $\phi(x)$ represents the *phase* of the wave. In a constant potential $\phi(x) = \pm k\,x$, that is, the phase changes linearly with x. In a slowly varying potential, it is expected that $\phi(x)$ deviates slightly from the linear dependence on x. To further investigate the function $\phi(x)$ it is convenient to use the abbreviations

$$k(x) = \frac{1}{\hbar}\sqrt{2m\left[E - U(x)\right]} \qquad \text{for} \qquad E \geq U(x) \tag{9.3}$$

$$k(x) = \frac{-\mathrm{i}}{\hbar}\sqrt{2m\left[U(x) - E\right]} = -\mathrm{i}\,\kappa(x) \qquad \text{for} \qquad E \leq U(x). \tag{9.4}$$

Insertion of Eq. (9.2) into the time-independent Schrödinger equation and using Eqs. (9.3) and (9.4) yields

$$\mathrm{i}\,\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} - \left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)^2 + k(x)^2 = 0 \tag{9.5}$$

which is just a different representation of the Schrödinger equation, which has, however, the same physical content. The WKB approximation is intended for potentials that do not vary rapidly. Therefore, as a zero-order approximation, we assume that the second derivative of $\phi(x)$ with respect to x is very small

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} \approx 0. \tag{9.6}$$

One obtains

$$\left(\frac{\mathrm{d}\phi_0}{\mathrm{d}x}\right)^2 = k(x)^2 \,. \tag{9.7}$$

The subscript zero in ϕ_0 is used to emphasize that this is a zero-order approximation. Integration yields

$$\boxed{\phi_0(x) = \pm\int k(x)\,\mathrm{d}x + C_0} \tag{9.8}$$

where C_0 is an integration constant. This equation is the simplest form of the WKB approximation.

A successive approximation method can be obtained by taking into account a *finite* second derivative, instead of the more crude approximation made in Eq. (9.6). Equation (9.5) can then be rewritten as

$$\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)^2 = k(x)^2 + \mathrm{i}\,\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} \,. \tag{9.9}$$

Integration of $\phi' = \mathrm{d}\phi/\mathrm{d}x$ without neglecting the second derivative on the right-hand side of this equation (in contrast to the previous omission of $\phi'' = \mathrm{d}^2\phi/\mathrm{d}x^2$, see Eq. 9.6) yields

$$\phi(x) = \pm\int_x \sqrt{k(x)^2 + \mathrm{i}\,\phi_0''(x)}\;\mathrm{d}x + C_1 \,. \tag{9.10}$$

Using Eq. (9.7) to determine $\phi''(x)$, one obtains

$$\phi_1(x) = \pm\int_x \sqrt{k(x)^2 + \mathrm{i}\,\phi_0''(x)}\;\mathrm{d}x + C_1 = \pm\int_x \sqrt{k(x)^2 \pm \mathrm{i}\,k'(x)}\;\mathrm{d}x + C_1 \tag{9.11}$$

where $k'(x) = \mathrm{d}k(x)/\mathrm{d}x$. The subscript one in ϕ_1 is used to emphasize that this is a first-order approximation. Since we have required that the wave function does not vary too violently, one can state that

$$\left|k'(x)\right| << \left|k(x)^2\right| \,. \tag{9.12}$$

Thus, the first-order approximation is only slightly different from the zero-order approximation of Eq. (9.8). With the condition of Eq. (9.12), one can further simplify Eq. (9.11) and expand the square-root as follows

$$\phi_1(x) \approx \int_x \left(\pm k(x) + \frac{\mathrm{i}}{2} \frac{k'(x)}{k(x)} \right) \mathrm{d}x + C_1 \approx \pm \int_x k(x)\,\mathrm{d}x + \frac{\mathrm{i}}{2} \ln k(x) + C_1 \,. \quad (9.13)$$

We have thus derived the zero-order (Eq. 9.8) and the first-order approximation (Eqs. 9.11 and 9.13) for $\phi(x)$. Both approximations are known as the WKB method. Insertion of Eqs. (9.8) and (9.13) into Eq. (9.2) yields the wave function for the zero order and first order WKB approximation

$$\boxed{\psi(x) \approx \exp\left(\pm \mathrm{i} \int_x k(x)\,\mathrm{d}x \right)} \qquad \text{(zero-order WKB)} \qquad (9.14)$$

$$\boxed{\psi(x) \approx \frac{1}{\sqrt{k(x)}} \exp\left(\pm \mathrm{i} \int_x k(x)\,\mathrm{d}x \right)} \qquad \text{(first-order WKB)} \qquad (9.15)$$

This wave function can be used in the classically allowed and forbidden regions, as illustrated by region II and III in ***Fig***. 9.1, respectively.

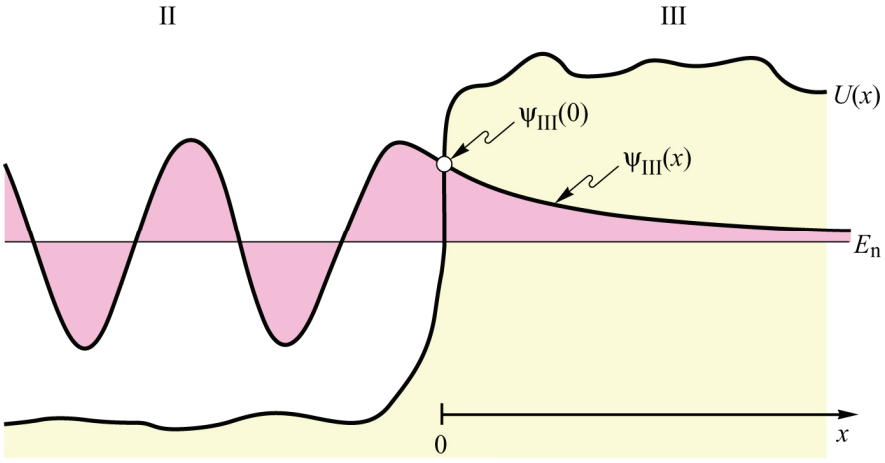


Fig. 9.1 Oscillating and exponentially decaying wavefunction in the classically allowed region II and disallowed region III, respectively. The decaying wave function in region III can be calculated by the WKB approximation.

As an example, we use the WKB approximation to calculate the amplitude of a wave function in the classically forbidden region of a potential energy $U(x)$. Such a classically forbidden region is shown in ***Fig***. 9.1. A particle with total energy E has a kinetic energy $E - U(x)$. In the classically forbidden region, the kinetic energy is negative which cannot occur for classical particles. Therefore, regions where the condition $E - U(x) < 0$ is satisfied, are forbidden for classical particles. In contrast, the Schrödinger equation has solutions even in classically forbidden regions. As will be seen below, the amplitude of the wave function rapidly decreases in these regions. The particle or the wave function ***tunnels*** into the barrier. The amplitude of the wave function in the barrier will be denoted as $\psi_{III}(x)$, as shown in ***Fig***. 9.1. To calculate $\psi_{III}(x)$, we assume that the wave function has an amplitude $\psi_{III}(0)$ at the boundary

between the classically allowed region II and the classically forbidden region III. Calculating the amplitude of the wave function according to the zero-order WKB approximation Eq. (9.14) yields

$$\psi_{III}(x) = \psi_{III}(0) \exp\left[-\int_0^x \kappa(x)\,dx\right] = \psi_{III}(0) \exp\left\{-\int_0^x \frac{1}{\hbar}\sqrt{2m\left[U(x)-E\right]}\,dx\right\} \quad (9.16)$$

Equations (9.14), (9.15), and (9.16) are valid for an arbitrary potential $U(x)$.

Applying the WKB approximation to the classically allowed region II, provides further insight. In region II, it is $E > U(x)$ and $k(x)$ is a real quantity. We define the effective wavelength as

$$\lambda(x) = \frac{2\pi}{k(x)} . \quad (9.17)$$

If $U(x)$ is constant, then $k(x)$ is a constant as well, and, according to Eq. (9.17), λ is a constant. In the classical picture, a particle in a constant potential has a constant momentum. In the quantum-mechanical picture, the wave in a constant potential has a constant wavelength. Now consider a varying potential energy $U(x)$. In this case, the momentum of a classical particle varies according to the variations of the potential energy. The wavelength of the quantum wave depends on the potential energy $U(x)$. This dependence is given by Eqs. (9.17) and (9.4). Hence, the WKB method can be understood as going from a *constant* wavelength λ in a *constant* potential to a *slowly varying* wavelength in a *slowly varying* potential. Recall that the WKB approximation was derived for a slowly varying potential (see Eq. 9.12). Using Eq. (9.17), this condition can be rewritten as

$$\frac{dU(x)}{dx} << \frac{U(x)}{\lambda} . \quad (9.18)$$

That is, the changes of $U(x)$ must be slow on the scale of λ.

Exercise 1: ***Tunneling probability through a barrier***. Consider the quantum barrier shown in ***Fig***. 9.2. Electrons in Region I do not have sufficient energy to overcome the barrier by thermal emission over the barrier. However, carriers have a non-zero probability to tunnel through the barrier. The probability of "finding" the electron at the left-hand side of the barrier is given by $\psi_{II}^*(0)\psi_{II}(0)$. Similarly, the probability of "finding" the electron at the right-hand side of the barrier is given by $\psi_{II}^*(L_B)\,\psi_{II}(L_B)$. Thus the tunneling probability is given by

$$T = \frac{\psi_{II}^*(L_B)\,\psi_{II}(L_B)}{\psi_{II}^*(0)\,\psi_{II}(0)} . \quad (9.19)$$

Using the zero-order WKB approximation to calculate the wave function in the barrier yields

$$\psi_{II}(x) = \psi_{II}(0)\,e^{-\int_0^x \kappa(x)dx} = \psi_{II}(0)\,e^{-\int_0^x \hbar^{-1}\sqrt{2m^*\left[U(x)-E\right]}\,dx} . \quad (9.20)$$

Thus the tunneling probability is given by

$$T = e^{-\int_{x=0}^{L_B} 2\hbar^{-1}\sqrt{2m^*[U(x)-E]}\,dx} \tag{9.21}$$

Calculate the tunneling probability of an electron with mass $m^* = 0.067m_0$ and energy $E = 100$ meV through a rectangular barrier with height $U_{II} = 200$ meV and barrier thickness $L_B = 100$ Å. Assume that $U_I = U_{III} = 0$. What is the dependence of the tunneling probability through a rectangular barrier on (*i*) the barrier thickness, (*ii*) the barrier height? What is the tunneling probability for $L_B = 1000$ Å?

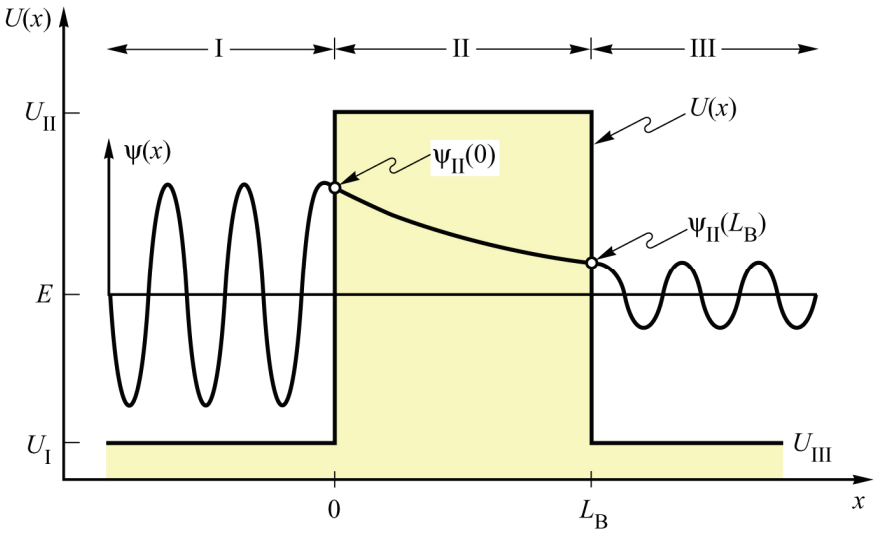


Fig. 9.2. Wave function of a particle with energy E tunneling through a quantum barrier.

Solution: The tunneling probability is $T = 2.2 \times 10^{-4}$ and $T = 2.8 \times 10^{-37}$ for $L_B = 100$ Å and 1000 Å, respectively. The tunneling probability decreases exponentially with the barrier thickness. It decreases exponentially with the square root of the barrier height. This exercise shows that tunneling effects are significant for barrier thicknesses on the order of 100 Å and smaller. It also shows that the tunneling probability through thick barriers (*e. g.* 1000 Å) is extremely small.

What are the implications for the gate leakage current in devices such as Si MOSFETs with an oxide thickness of 100 Å?

Solution: In Si MOSFETs, the tunneling barrier height is >> 100 meV since SiO_2 has a large bandgap ($E_g \approx 5$ eV). Therefore, the gate leakage current is negligibly small for an oxide thickness of 100 Å.

What are the implications for the doping concentration in ohmic metal-semiconductor contacts?

Solution: Ohmic contacts are made by heavily doping the semiconductor of a metal-semiconductor contact. The depletion region thickness in the semiconductor is so small that tunneling is the dominant transport mechanism between the metal and the semiconductor. Thus, the doping concentration in the semiconductor must be so high that the depletion region thickness is << 100 Å.

9.2 The connection formulas in the WKB method

The WKB approximation provides a quasi-oscillatory solution and a quasi-exponentially damped solution in the classically allowed and forbidden regions of a potential respectively. Special care must be taken to use the WKB method in the vicinity of a so-called classical turning point. As shown in ***Fig***. 9.3, a classical particle would be allowed only in region II where $E > U(x)$. A classical particle with total energy E would *turn around* at $x = a$, since its kinetic energy at this point is zero, *i. e*. $E - U(a) = 0$. In contrast, a quantum-mechanical particle can tunnel beyond the classical turning point.

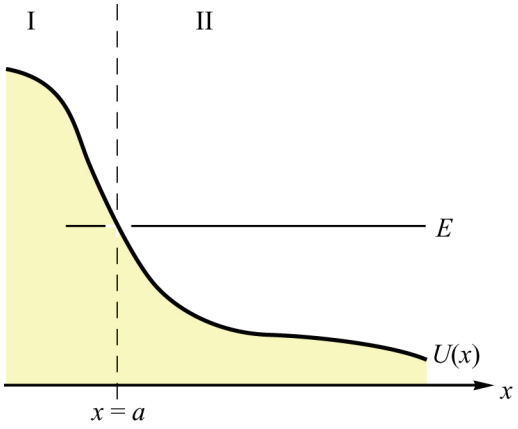


Fig. 9.3. The classical turning point at x = a is to the left of the classically allowed region II.

In the vicinity of the classical turning point, the wavevector approaches zero, *i. e*. $k(x) \to 0$. However, the derivative remains finite, that is

$$\frac{\mathrm{d}k(x)}{\mathrm{d}x} \quad = \quad \frac{\mathrm{d}}{\mathrm{d}x}\frac{-\mathrm{i}}{\hbar}\sqrt{2\,m\left[E-U(x)\right]} \quad \neq \quad 0\ . \tag{9.22}$$

The application of the WKB method requires that $k(x)$ does note vary violently, as stated in Eqs. (9.12) and (9.18). This condition can be also expressed as

$$\Delta\, k(x) \quad = \quad \frac{\mathrm{d}k(x)}{\mathrm{d}x}\ \Delta x \quad << \quad k(x)\ . \tag{9.23}$$

This condition is obviously not fulfilled in the vicinity of a classical turning point where $k(x) \to 0$. Hence, the WKB approximation cannot be applied in the vicinity of classical turning points. However, because the WKB method is problematic *only* in the vicinity of a turning point, it is desirable to find some *connection formulas* that would make possible the utilization of the WKB approximation even in the vicinity of turning points.

An excellent derivation of the connection formulas was given by Merzbacher (1970). In this derivation, it was assumed that the potential energy $U(x)$ depends linearly on x in the vicinity of the classical turning point. The derivation of the connection formulas shall be omitted here and only the results will be summarized. The connection formulas *connect* the wave functions obtained by the WKB-method in regions I, II, and III (see ***Fig***. 9.3 and ***Fig***. 9.4). The classical turning point may be to the left or to the right of the allowed region II as shown in ***Fig***. 9.3 and ***Fig***. 9.4, respectively. The connection formulas are given by:

(1) Turning point is to the left of the classically allowed region (see ***Fig***. 9.3)

$$\frac{1}{\sqrt{\kappa}}\,\exp\left(-\int_x^a \kappa\,dx\right) \qquad \text{connects with} \qquad \frac{2}{\sqrt{k}}\,\cos\left(\int_a^x k\,dx - \frac{\pi}{4}\right) \tag{9.24}$$

$$\frac{-1}{\sqrt{\kappa}}\,\exp\left(\int_x^a \kappa\,dx\right) \qquad \text{connects with} \qquad \frac{1}{\sqrt{k}}\,\sin\left(\int_a^x k\,dx - \frac{\pi}{4}\right) \tag{9.25}$$

(2) Turning point is to the right of the classically allowed region (see ***Fig***. 9.3)

$$\frac{2}{\sqrt{k}}\,\cos\left(\int_x^b k\,dx - \frac{\pi}{4}\right) \qquad \text{connects with} \qquad \frac{1}{\sqrt{\kappa}}\,\exp\left(-\int_b^x \kappa\,dx\right) \tag{9.26}$$

$$\frac{1}{\sqrt{k}}\,\sin\left(\int_x^b k\,dx - \frac{\pi}{4}\right) \qquad \text{connects with} \qquad \frac{-1}{\sqrt{\kappa}}\,\exp\left(\int_b^x \kappa\,dx\right) \tag{9.27}$$

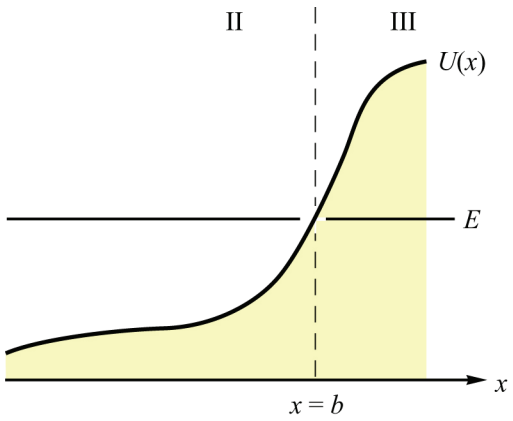


Fig. 9.4. The classical turning point at x = b is to the right of the classically allowed region II.

Caution has to be exercised, if the connection formulas are applied. Consider Case (1), in which the classical turning point is to the left of the classically allowed region. According to Eqs. (9.24) and (9.25), the wave function on the left side of the turning point may either *decrease* exponentially (Eq. 9.25) or *increase* exponentially (Eq. 9.24) with x. Even though a wave function may be adequately described by a single exponential function at some position x_0, the connecting wave function may become important at another position x_1, due to the exponential nature of the two functions. Therefore, caution must be exercised, if one of the two wave functions is neglected.

9.3 The WKB method for bound states

The WKB method can be used to obtain the eigenstate energies of a potential well. An example of such a potential well is shown in ***Fig***. 9.5. The WKB approximation can be used in the three regions I, II, and III. In the vicinity of the classical turning points, the connection formulas will be used. In region I, the wave function must vanish for sufficiently small x. Therefore, the unnormalized wave function in region I is given by

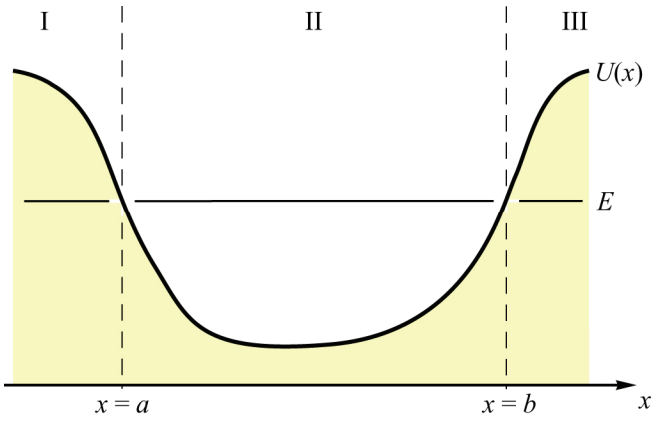


Fig. 9.5. Schematic potential well with potential energy $U(x)$ and two classical turning points $x = a$ and $x = b$. A classical particle of energy E would be confined to the region $a \le x \le b$.

$$\psi_I \approx \frac{1}{\sqrt{\kappa}} \exp\left(-\int_x^a \kappa \, dx\right) \qquad (x < a). \tag{9.28}$$

According to the connection formula Eq. (9.24) the wave function in the classically allowed region II is given by

$$\psi_{II} \approx \frac{2}{\sqrt{k}} \cos\left(\int_a^x k \, dx - \frac{\pi}{4}\right) \qquad (a \le x \le b). \tag{9.29}$$

Rewriting this equation by employing a trigonometric conversion yields

$$\psi_{II} \approx \frac{2}{\sqrt{k}} \cos\left(\int_a^b k \, dx - \int_x^b k \, dx - \frac{\pi}{4}\right) \tag{9.30}$$

$$= \frac{-2}{\sqrt{k}} \cos\left(\int_a^b k \, dx\right) \sin\left(\int_x^b k \, dx - \frac{\pi}{4}\right) + \frac{2}{\sqrt{k}} \sin\left(\int_a^b k \, dx\right) \cos\left(\int_x^b k \, dx - v\frac{\pi}{4}\right).$$

Comparing the two terms of this equation with the connection formulas, Eqs. (9.26) and (9.27), yields that only the last cosine term yields an exponentially decreasing wave function in region III. Hence, the first term must vanish; this yields the condition

$$\int_a^b k(x) \, dx = \left(n + \tfrac{1}{2}\right)\pi \qquad \text{for} \qquad n = 0, 1, 2 \ldots \tag{9.31}$$

This equation enables us to obtain the discrete (approximate) eigenstate energies of an arbitrary shaped quantum well. Note that the validity of the connection formulas is limited to potential energies $U(x)$ which depend linearly on x in the vicinity of the classical turning point. With the de Broglie relation $p = \hbar k$ one obtains

$$\int_a^b p(x) \, dx = \left(n + \tfrac{1}{2}\right)\pi\hbar \qquad \text{for} \qquad n = 0, 1, 2 \ldots \tag{9.32}$$

Integration of $p(x)$ in terms of a closed curve (*i. e.* one round trip of the quantum wave in the potential well) yields

$$\oint p(x)\,dx \;=\; 2\int_a^b p(x)\,dx \;=\; \left(n + \tfrac{1}{2}\right) 2\pi\hbar \qquad \text{for} \qquad n = 0, 1, 2 \ldots \tag{9.33}$$

The physical interpretation of this equation is facilitated by rewriting the equation as

$$\boxed{\frac{2}{\hbar}\int_a^b p(x)\,dx \;=\; 2\pi n + 2\pi c_a + 2\pi c_b} \qquad \text{for} \qquad n = 0, 1, 2 \ldots \tag{9.34}$$

The term $2\pi n$ represents the integral number of wavelengths of the quantum mechanical wave in the quantum well between the classical turning points. In the example shown in ***Fig***. 9.6, three full wavelengths fit into the round-trip distance of the quantum well and hence $n = 3$. ***The terms $2\pi c_a$ and $2\pi c_b$ in Eq. (9.34) can be interpreted as the change of the phase of the quantum mechanical wave incurred at the classical turning points $x = a$ and $x = b$***. Comparison of Eq. (9.34) with Eq. (9.33) reveals that $c_a = c_b = (1/4)$, *i. e.*, the change of phase of the wave function at the classical turning points is a quarter wave. Note that this change of phase of $\pi/2$ deduced here is the result of the connection formulas discussed in the previous section. The connection formulas were derived for potential energies $U(x)$ that are, in the vicinity of the turning points, linear functions of x. This situation is shown in ***Fig***. 9.6(a). Consequently, a phase change of $\pi/2$ applies only to potentials $U(x)$ that depend linearly on x in the vicinity of the turning point.

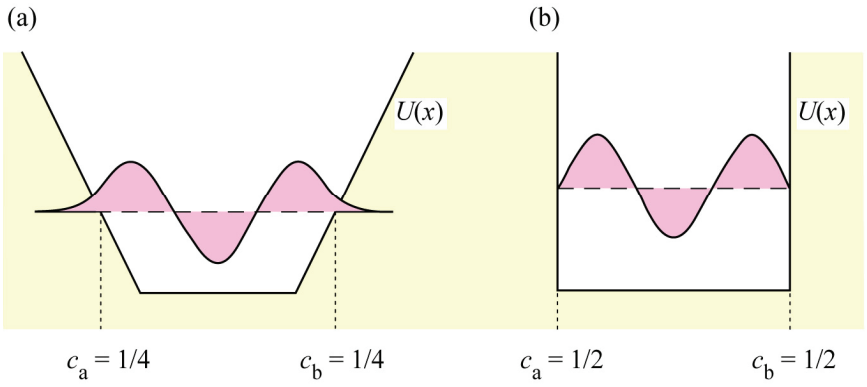


Fig. 9.6. Phase changes $c_a 2\pi$ and $c_b 2\pi$ of wave functions in quantum wells for a potential energy $U(x)$ varying (a) linearly and (b) discontinuously at the classical turning points.

We now consider the case in which the turning points are at a *discontinuous change* of $U(x)$, as illustrated in ***Fig***. 9.6(b). Assuming that the walls of the potential well are infinitely high, the wave function must vanish at the turning point. Hence the change in phase incurred by the wave at the turning point must be π, *i. e.* $c_a = c_b = (1/2)$. Using this value in Eq. (9.34), one obtains

$$\int_a^b p(x)\,dx \;=\; (n + 1)\,\pi\hbar \qquad \text{for} \qquad n = 0, 1, 2 \ldots \tag{9.35}$$

which is identical to the one-dimensional Bohr-Sommerfeld quantization condition given in Eq. (1.3.36). This is not surprising: In the Bohr-Sommerfeld model, the electron trajectory around the nucleus is a rigid circle or ellipsis, and hence the wave function vanishes for radii larger than the Bohr radius.

We summarize the phase changes of the quantum mechanical wave for a linear and a discontinuous potential energy $U(x)$ at the turning point:

$$\text{If} \quad U(x) = \text{linear function, then} \quad c_a = \frac{1}{4} \tag{9.36}$$

$$\text{If} \quad U(x) = \text{discontinous, then} \quad c_a = \frac{1}{2}\,. \tag{9.37}$$

The same applies to c_b.

Exercise 2: ***The WKB approximation for bound states***. Calculate the energies of the allowed states of a particle with mass m^* in a one-dimensional ***infinite square-shaped*** quantum well by using the WKB approximation. The potential energy of the infinite square well is given by $U(x) = 0$ for $|x| < L_{QW}/2$ and $U(x) = \infty$ for $x = \pm L_{QW}/2$.

Solution: Using Eq. (9.34) with $c_a = c_b = (1/2)$, one obtains

$$\int_{-L_{QW}/2}^{L_{QW}/2} p(x)\,dx = \int_{-L_{QW}/2}^{L_{QW}/2} \hbar \frac{\sqrt{2m^*\left(E_n - U(x)\right)}}{\hbar}\,dx = \int_{-L_{QW}/2}^{L_{QW}/2} \sqrt{2m^*\left(E_n - 0\right)}\,dx = \tag{9.38a}$$

$$= L_{QW}\sqrt{2m^* E_n} = (n+1)\pi\hbar \tag{9.38b}$$

Solving for E_n yields

$$E_n = (n+1)^2 \frac{\pi^2 \hbar^2}{2\,m^* L_{QW}^2} \qquad \text{for } n = 0, 1, 2 \ldots \tag{9.38c}$$

Calculate the energies of the allowed states of a particle with mass m^* in a one-dimensional ***triangular-shaped*** quantum well by using the WKB approximation. The potential energy of the triangular-shaped well is given by $U(x) = e\mathcal{E}x$ for $x \geq 0$ and $U(x) = \infty$ for $x < 0$.

Solution: Using Eq. (9.34) with $c_a = (1/2)$ and $c_b = (1/4)$, one obtains

$$E_n = \left(\frac{3\pi}{2}\left(n + \frac{3}{4}\right)\right)^{2/3} \left(\frac{e^2 \hbar^2 \mathcal{E}^2}{2\,m^*}\right)^{1/3} \qquad \text{for } n = 0, 1, 2 \ldots \tag{9.39}$$

Calculate the energies of the allowed states of a particle with mass m^* in a one-dimensional ***V-shaped*** quantum well by using the WKB approximation. The potential energy of the V-shaped well is given by $U(x) = e\mathcal{E}|x|$.

Solution: $$E_n = \left(\frac{3\pi}{4}\left(n + \frac{1}{2}\right)\right)^{2/3} \left(\frac{e^2 \hbar^2 \mathcal{E}^2}{2\,m^*}\right)^{1/3} \qquad \text{for } n = 0, 1, 2 \ldots \tag{9.40}$$

Calculate the energies of the allowed states of a particle with mass m^* in a one-dimensional ***parabolic-shaped*** quantum well by using the WKB approximation. The potential energy of the parabolic-shaped well is given by $U(x) = (1/2)\,m\,\omega^2 x^2$.

Solution: $E_n = \left(n + \frac{1}{2}\right)\hbar\,\omega$ for $n = 0, 1, 2 \ldots$ (9.41)

9.4 The variational method

The variational method is a versatile method to calculate approximate wave functions and eigenstate energies of a potential. The starting point of the variational method is an educated guess for the wave function of a given quantum mechanical potential. Having made such an initial guess, it is not clear that the wave function given by the educated guess is a *good* wave function, *i. e*., if it matches well the true wave function of the potential. Therefore, we call the approximate wave function based on the guess the trial wave function or ***trial function***. The trial function may contain one or several trial parameters whose selection allows for an optimization of the wave function. Assuming that the normalized trial function for a one-dimensional problem is then given by $\psi(x, \alpha)$, where x is the spatial coordinate and α is the trial function parameter or ***trial parameter***, whose value must yet be determined.

The expectation value for the total energy of a quantum mechanical particle, described by the trial function $\psi\,(x, \alpha)$, is given by

$$\langle E \rangle = \langle \psi(x, \alpha) \,|H|\, \psi(x, \alpha) \rangle \tag{9.42}$$

Like ordinary particles, quantum mechanical particles assume a state in which their total energy is minimized. This fact can be used to find an optimum value for the trial parameter. Thus the condition

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\langle E \rangle = \frac{\mathrm{d}}{\mathrm{d}\alpha}\langle \psi(x, \alpha) \,|H|\, \psi(x, \alpha) \rangle = 0 \tag{9.43}$$

can be used to determine the optimum value for α. Of course the quality of the wave function and of the eigenstate energy will also depend on the quality of the trial function, *i. e*., on the initial educated guess. Finally, the insertion of the optimum value for α into Eq. (9.42) yields the eigenstate energy.

If the trial function has more than one trial parameter, the condition of Eq. (9.43) needs to be applied several times, *i. e*. for each trial parameter. Whereas the quality of the wave function can improve with the number of trial parameters, so does the computational effort. The variational method thus allows one to calculate the wave function and the eigenstate energy of particles in a quantum mechanical potential. With a good trial function, the accuracy of the eigenstate energy can be better than 1%, even if only a single trial parameter is used.

Exercise 3: ***The Fang-Howard wave function***. In this exercise, the ground-state wave function and energy of a triangular potential well are calculated by the variational method. The potential occurring in the semiconductor of a MIS structure has a triangular shape, as shown in ***Fig***. 9.7. The potential energy in the semiconductor is given by $U(x) = E_{\mathrm{pot}}(x) = e\,\mathcal{E}\,x$. Use the trial function $\psi(x, \alpha) = A\,x\,\mathrm{e}^{-\alpha x}$ where A is a normalization constant and α is the trial parameter. This trial function was introduced by Fang and Howard (1966).

First calculate the normalization constant A by using the normalization condition. Then calculate the expectation value of the total energy as a function of α. Determine the trial

parameter α by minimizing the expectation value of the total energy with respect to α, *i. e.* $(d/d\alpha)\langle E(\alpha)\rangle = 0$. Finally calculate the expectation value for the total energy, E_0.

Solution: Application of the normalization condition (*i. e.* using $\langle\psi|\psi\rangle = 1$) yields:

$$A = 2\alpha^{3/2} \tag{9.44a}$$

Minimizing the energy expectation value (*i. e.* using $d\langle E\rangle/d\alpha = 0$) yields:

$$\alpha = [(3/2)\, e\, \mathcal{E}\, m^*/\hbar^2]^{1/3} \tag{9.44b}$$

Calculation of the total energy expectation value (*i. e.* calculating $\langle E\rangle$) yields:

$$E_0 = (3/2)[(3/2)\, e\, \mathcal{E}\, \hbar/(m^*)^{1/2}]^{2/3} \tag{9.44c}$$

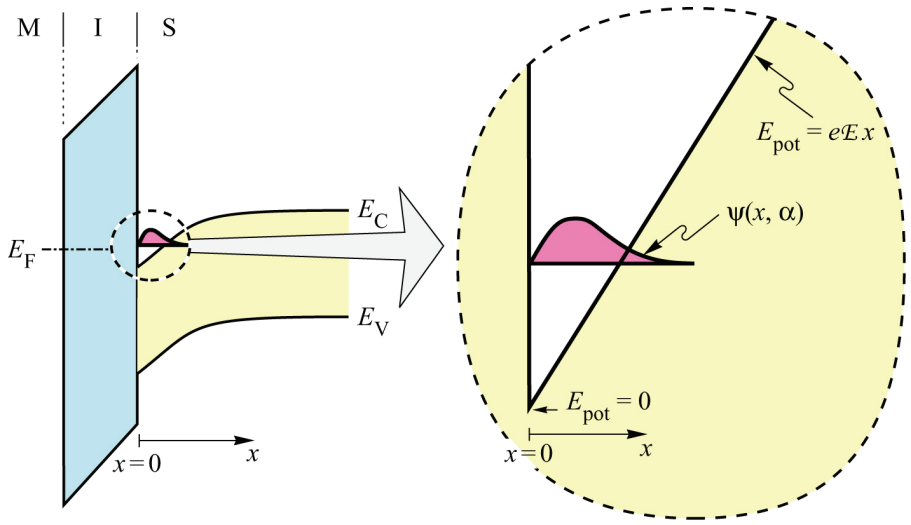


Fig. 9.7. Band diagram of an triangular-shaped electron channel of a metal insulator-semiconductor structure. Also shown is the Fang-Howard wave function.

Draw the band diagram of a GaAs metal insulator semiconductor (MIS) structure for negative, zero, and positive bias. Then assume that the electric field in the GaAs is $\mathcal{E} = 7 \times 10^4$ V/cm in the positively-biased case. Calculate the ground state energy of electrons in the GaAs ($m^* = 0.067\, m_0$) by using the Fang-Howard wave function.

Solution: $E_0 = 75$ meV.

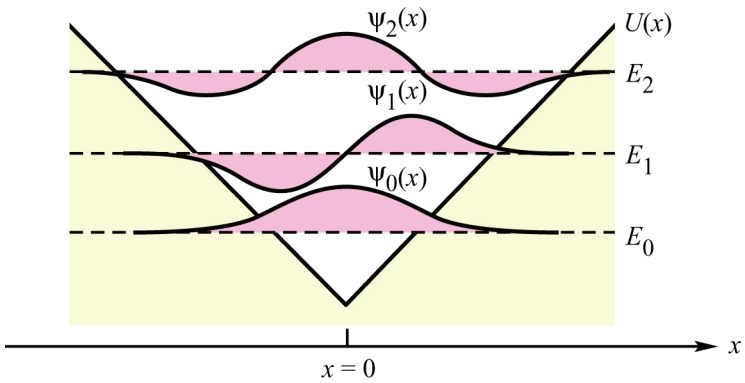


Fig. 9.8. Potential energy diagram and schematic wave functions of a V-shaped quantum well.

Exercise 4: ***Variational wave functions in a V-shaped potential well***. In this exercise, the ground-state wave function and two excited-state wave functions and the state energies of a V-shaped potential well are calculated by the variational method. The potential and wave functions are shown in ***Fig***. 9.8. The potential energy in the semiconductor is given by $U(x) = E_{pot}(x) = e\,\mathcal{E}\,x$ for $x \geq 0$ and $U(x) = E_{pot}(x) = -e\,\mathcal{E}\,x$ for $x \leq 0$. Develop trial wave functions of the lowest three states taking into account that the lowest three wave functions have 0, 1, and 2 nodes. Also take into account the *even* symmetry of the $n = 0$ and $n = 2$ states, and the *odd* symmetry of the $n = 1$ state.

Solution: There are several possible solutions to this exercise. One possible set of trial wave functions is:

$$\psi_0(x) = A_0\,(1 + \alpha_0\,x)\,e^{-\alpha_0 x} \quad \text{for} \quad x \geq 0 \tag{9.45a}$$

$$\psi_0(x) = A_0\,(1 - \alpha_0\,x)\,e^{\alpha_0 x} \quad \text{for} \quad x < 0 \tag{9.45b}$$

$$\psi_1(x) = A_1\,x\,e^{-\alpha_1 x} \quad \text{for} \quad x \geq 0 \tag{9.45c}$$

$$\psi_1(x) = A_1\,x\,e^{\alpha_1 x} \quad \text{for} \quad x < 0 \tag{9.45d}$$

$$\psi_2(x) = A_2\,(\alpha_2^2\,x^2 - 1)\,(1 + \alpha_2\,x)\,e^{-\alpha_2 x} \quad \text{for} \quad x \geq 0 \tag{9.45e}$$

$$\psi_2(x) = A_2\,(\alpha_2^2\,x^2 - 1)\,(1 - \alpha_2\,x)\,e^{\alpha_2 x} \quad \text{for} \quad x < 0 \tag{9.45f}$$

The normalization constants of the trial wave functions are determined by using the normalization condition. One obtains

$$A_0 = \sqrt{(2/5)\,\alpha_0} \tag{9.46a}$$

$$A_1 = \sqrt{2\,\alpha_1^3} \tag{9.46b}$$

$$A_2 = \sqrt{(4/63)\,\alpha_2} \tag{9.46c}$$

Minimizing the expectation value of the energy yields the variational parameters

$$\alpha_0 = \left(\frac{9}{4}\right)^{1/3} \left(e\,\mathcal{E}\,2\,m^*/\hbar^2\right)^{1/3} \tag{9.47a}$$

$$\alpha_1 = \left(\frac{3}{4}\right)^{1/3} \left(e\,\mathcal{E}\,2\,m^*/\hbar^2\right)^{1/3} \tag{9.47b}$$

$$\alpha_2 = \left(\frac{47}{12}\right)^{1/3} \left(e\,\mathcal{E}\,2\,m^*/\hbar^2\right)^{1/3} \tag{9.47c}$$

The eigenstate energies are then given by

$$E_0 = \frac{3}{10}\left(\frac{81}{2}\right)^{1/3}\left(\frac{e^2\,\hbar^2\,\mathcal{E}^2}{2\,m^*}\right)^{1/3} \tag{9.48a}$$

$$E_1 = \frac{3}{2}\left(\frac{9}{2}\right)^{1/3}\left(\frac{e^2\,\hbar^2\,\mathcal{E}^2}{2\,m^*}\right)^{1/3} \tag{9.48b}$$

$$E_2 = \frac{9}{7}\left(\frac{47}{12}\right)^{2/3}\left(\frac{e^2\,\hbar^2\,\mathcal{E}^2}{2\,m^*}\right)^{1/3} \tag{9.48c}$$

The agreement of the eigenstate energies with the numerically evaluated correct values is generally quite good. The difference between approximation and correct value is typically just a few percent.

References

Bohm D. *Quantum Theory* (Prentice-Hall, Englewood Cliffs, 1951)

Fang F. F. and Howard, W. E. "Negative field-effect mobility on (100) Si surfaces" *Physical Review Letters*, **16**, 797 (1966)

Wentzel G., Kramers H. A., and Brillouin L. The WKB approximation was developed independently and simultaneously by the three authors (1926)

10

Perturbation theory

Quantum mechanical systems may be exposed to perturbations including external electric fields, magnetic fields, or electromagnetic radiation. Due to such perturbations, the quantum system considered here is stimulated and, as a consequence, changes its state. This change of state may include changes in the shape of wave functions, state energies, and occupation probability of states. This is what perturbation theory is all about. Perturbation theory is one of the most important methods for obtaining approximate solutions to Schrödinger's equation.

10.1 First-order time-independent perturbation theory

This section covers first-order perturbation calculation of a stationary, non-degenerate quantum state. Suppose a quantum mechanical system whose eigenstate energies and wave functions are known. Suppose that the unperturbed system is described by the hamiltonian operator H^0, the eigenstate energies E_n^0, and the wave functions ψ_n^0. Then the Schrödinger equation of the unperturbed system is given by

$$H^0 \, \psi_n^0 \quad = \quad E_n^0 \, \psi_n^0 \, . \tag{10.1}$$

Here, the superscript 0 is used for energies, wave functions and the hamiltonian operator of the *unperturbed* system. If the system is subjected to a small perturbation, then perturbation theory allows one to determine the modifications of the eigenstate energies, wave functions, and occupation probabilities. It may seem that these are very special circumstances; however, it will become clear, that perturbation theory is of great practical importance.

The hamiltonian operator of a perturbed system is given by

$$H \quad = \quad H^0 \, + \, \lambda \, H' \tag{10.2}$$

where H^0 is the hamiltonian operator of the unperturbed system and H' is called the ***perturbation term*** in the hamiltonian. The parameter λ allows us to turn the perturbation on ($\lambda = 1$) and off ($\lambda = 0$). The parameter λ further indicates the smallness of the perturbation. That is, the system described by the hamiltonian H has experienced only a *small* perturbation when compared to the unperturbed system. The parameter λ can have the value of one, ($\lambda = 1$), without loss of general validity of the perturbation theory. The Schrödinger equation of the perturbed system is given by

$$H \, \psi_n \quad = \quad \left(H^0 \, + \, \lambda \, H' \right) \psi_n \quad = \quad E_n \, \psi_n \, . \tag{10.3}$$

It is evident that the perturbed system merges with the unperturbed system if λ approaches zero, *i. e.*

$$\lim_{\lambda \to 0} E_n \quad = \quad E_n^0 \tag{10.4}$$

and

$$\lim_{\lambda \to 0} \psi_n \;=\; \psi_n^0 \,. \tag{10.5}$$

To obtain a solution of the perturbed problem, an expansion of E_n and ψ_n in a power series in λ is employed

$$\boxed{E_n \;=\; E_n^0 \;+\; \lambda\, E_n' \;+\; \lambda^2\, E_n'' \;+\; \cdots} \tag{10.6}$$

$$\boxed{\psi_n \;=\; \psi_n^0 \;+\; \lambda\, \psi_n' \;+\; \lambda^2\, \psi_n'' \;+\; \cdots} \tag{10.7}$$

where $E_n' = \mathrm{d}E_n / \mathrm{d}\lambda$ and $\psi_n' = \mathrm{d}\psi_n / \mathrm{d}\lambda$. The first three terms of this power series are illustrated in ***Fig***. 10.1, where the unperturbed values of E_n and ψ_n are displayed together with their first-order correction term ($\lambda E_n'$ and $\lambda\psi_n'$) and their second-order correction term ($\lambda^2 E_n''$ and $\lambda^2\psi_n''$). It is the goal of ***first-order perturbation theory*** to find the values of E_n' and ψ_n'. Correspondingly, it is the aim of ***second-order perturbation theory*** to find the values of E_n'' and ψ_n''.

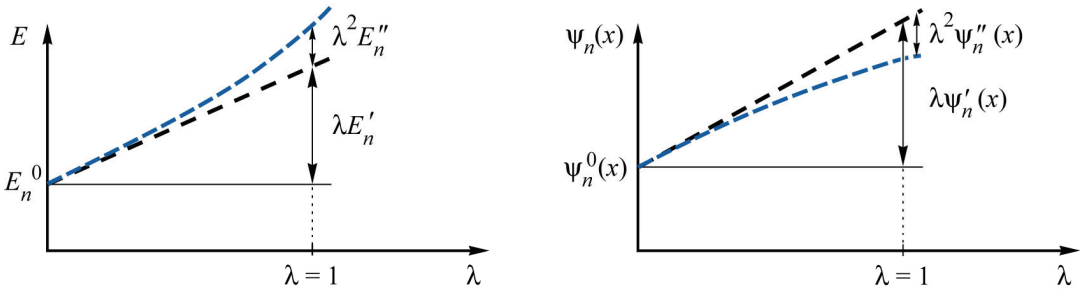


Fig. 10.1. Illustration of first-order and second-order corrections terms to the unperturbed energy E_n^0 and to the amplitude of the unperturbed wavefunction ψ_n^0 at a particular position x. The parameter λ is a parameter that controls the magnitude of the perturbation; $\lambda = 0$ corresponds to "no perturbation".

Substitution of Eqs. (10.6) and (10.7) into the perturbed Schrödinger equation (Eq. 10.3) yields

$$\begin{aligned}
&\lambda^0 \left(H^0\, \psi_n^0 \;-\; E_n^0\, \psi_n^0 \right) \\
+\; &\lambda^1 \left(H^0\, \psi_n' \;+\; H'\, \psi_n^0 \;-\; E_n^0\, \psi_n' \;-\; E_n'\, \psi_n^0 \right) \\
+\; &\lambda^2 \left(H^0\, \psi_n'' \;+\; H'\, \psi_n' \;-\; E_n^0\, \psi_n'' \;-\; E_n'\, \psi_n' \;-\; E_n''\, \psi_n^0 \right) \\
+\; &\lambda^3 \left(\cdots \right) \\
+\; &\lambda^4 \left(\cdots \right) \\
+\; &\cdots \;=\; 0 \,.
\end{aligned} \tag{10.8}$$

If the perturbation is neglected ($\lambda = 0$), one obtains the original eigenvalues E_n^0 and eigenfunctions ψ_n^0. We next consider the case $\lambda \neq 0$. The sum in Eq. (10.8) equals zero, only if each summand is zero, that is

0th order terms: $$H^0 \psi_n^0 = E_n^0 \psi_n^0 \tag{10.9}$$

1st order terms: $$H^0 \psi_n' + H' \psi_n^0 = E_n^0 \psi_n' + E_n' \psi_n^0 \tag{10.10}$$

2nd order terms: $$H^0 \psi_n'' + H' \psi_n' = E_n^0 \psi_n'' + E_n' \psi_n' + E_n'' \psi_n^0 \tag{10.11}$$

The first of these three equation is the unperturbed Schrödinger equation. The second equation contains only first-order terms and it will be used to derive first-order perturbation results. The third equation contains only first-order and second-order terms and it will be used to derive second-order perturbation results. Note that in the above three equations, H^0, ψ_n^0, E_n^0, and H' are known.

It is the purpose of first-order perturbation theory to find solutions for E_n' and ψ_n' (E_n'' and ψ_n'' will be determined by second-order perturbation theory which is discussed in the subsequent section). Because ψ_n' is an unknown wave function, we will try to *express* ψ_n' *as a series of a complete set of orthogonal eigenfunctions*

$$\psi_n' = \sum_j a_j \psi_j^0 \tag{10.12}$$

The wave functions ψ_j^0 represent the complete orthonormal set of wave functions of the system. Nearly any wave function can be synthesized from this orthonormal set. The particular wave function ψ_n is one specific wave function of the complete set. In order to determine the perturbed wave function ψ_n', the coefficients a_j must be determined. Substitution of Eq. (10.12) into Eq. (10.10) yields

$$H^0 \sum_j a_j \psi_j^0 + H' \psi_n^0 = E_n^0 \sum_j a_j \psi_j^0 + E_n' \psi_n^0 \,. \tag{10.13}$$

Using

$$H^0 \sum_j a_j \psi_j^0 = \sum_j a_j E_j^0 \psi_j^0 \tag{10.14}$$

one obtains

$$\sum_j a_j E_j^0 \psi_j^0 + H' \psi_n^0 = E_n^0 \sum_j a_j \psi_j^0 + E_n' \psi_n^0 \,. \tag{10.15}$$

Consider next a wave function ψ_m^0 which is the mth wave function of the orthonormal set of wave functions ψ_j^0 given in Eq. (10.12). Pre-multiplication of Eq. (10.15) with ψ_m^{0*}, and integration over position space yields

$$a_m E_m^0 + \left\langle \psi_m^0 \middle| H' \middle| \psi_n^0 \right\rangle = a_m E_n^0 + E_n' \delta_{mn} \,. \tag{10.16}$$

Here, we have used the orthogonality of the set ψ_j^0, that is $\langle \psi_m^0 | \psi_j^0 \rangle = 0$ for $m \neq j$ and

$\langle \psi_m{}^0 | \psi_j{}^0 \rangle = 1$ for $m = j$. This can be expressed by the Kronecker delta $\langle \psi_m{}^0 | \psi_j{}^0 \rangle = \delta_{mj}$ or $\langle \psi_m{}^0 | \psi_n{}^0 \rangle = \delta_{mn}$. In Eq. (10.16), it is either $m = n$ or $m \neq n$. One obtains for

$m = n$:
$$E_n' = \left\langle \psi_n^0 \middle| H' \middle| \psi_n^0 \right\rangle \tag{10.17}$$

which is the *first-order correction term* to the energy. Hence, the energy of the nth state of a system, subjected to the perturbation hamiltonian H', calculated by first-order perturbation theory, is given by

$$\boxed{E_n = E_n^0 + \lambda \left\langle \psi_n^0 \middle| H' \middle| \psi_n^0 \right\rangle} \tag{10.18}$$

Furthermore, one obtains for

$m \neq n$:
$$a_m = \frac{\left\langle \psi_m^0 \middle| H' \middle| \psi_n^0 \right\rangle}{E_n^0 - E_m^0} \tag{10.19}$$

This equation can be used to calculate all a_m, except the value of $a_{m=n}$. The value of $a_{m=n}$ can be calculated by requiring that the first-order corrected wave function is normalized, *i. e.* $\langle \psi_n{}^0 + \lambda \psi_n{}' | \psi_n{}^0 + \lambda \psi_n{}' \rangle = 1$. This condition yields for

$m = n$:
$$\int_{-\infty}^{\infty} \left(\psi_n^0 + \lambda \sum_m a_m \psi_m^0 \right)^* \left(\psi_n^0 + \lambda \sum_m a_m \psi_m^0 \right) dx = 1 + \lambda a_m + \lambda a_m^* + \lambda^2 \sum_m a_m a_m^* = 1 \tag{10.20}$$

A simple solution of this equation is $a_m = a_m^* = a_n = a_n^* = 0$. Hence, the wave function of the nth state of a system, subjected to the perturbation hamiltonian H', calculated by first-order perturbation theory, is given by

$$\boxed{\psi_n = \psi_n^0 + \lambda \sum_{m \neq n} \frac{\left\langle \psi_m^0 \middle| H' \middle| \psi_n^0 \right\rangle}{E_n^0 - E_m^0} \psi_m^0} \tag{10.21}$$

The sum in this equation is carried out for all values of m *except* the value $m = n$. This equation shows, that the wave functions of all other (unperturbed) states have to be known, to calculate the perturbed wave function of the nth state. The influence of other wave functions decreases, as the energy-separation increases, since $E_n{}^0 - E_m{}^0$ is in the denominator of the expression.

10.2 Second-order time-independent perturbation theory

The energy and the wave function of a perturbed state can be expressed in terms of the expansion of Eqs. (10.6) and (10.7). It is the aim of the *second-order* perturbation calculation to find formulas for E_n'' and ψ_n''. In analogy to the first-order calculation, ψ_n'' is expressed in terms of the complete set of orthonormal wave functions of the unperturbed system, *i. e.*

$$\boxed{\psi''_n \;=\; \sum_j b_j\,\psi_j^0} \tag{10.22}$$

Inserting Eq. (10.22) and Eq. (10.12) into Eq. (10.11) yields

$$\sum_j b_j E_j^0 \psi_j^0 \;+\; H' \sum_j a_j \psi_j^0 \;=\; \sum_j b_j E_n^0 \psi_j^0 \;+\; \sum_j a_j E'_n \psi'_j \;+\; E''_n \psi_n^0 \;. \tag{10.23}$$

Similar to the previous section, we consider one specific wave function $\psi_m{}^0$ of the complete orthonormal set of wave functions of the unperturbed system. Pre-multiplication with $\psi_m{}^{0*}$, integration over all configuration space, and recalling that $\langle \psi_m{}^0 | \psi_j{}^0 \rangle = \delta_{mj}$ yields

$$b_m E_m^0 \;+\; \sum_j a_j \left\langle \psi_m^0 \middle| H' \middle| \psi_j^0 \right\rangle \;=\; b_m E_n^0 \;+\; E'_n a_m \;+\; E''_n \delta_{nm} \;. \tag{10.24}$$

In this equation, it is either $m = n$ or $m \neq n$. With $m = n$, one obtains the *second-order correction term* to the energy

$m = n$:
$$E''_n \;=\; \sum_j a_j \left\langle \psi_n^0 \middle| H' \middle| \psi_j^0 \right\rangle \;-\; E'_n a_n \tag{10.25}$$

$$\;=\; \sum_{j\neq n} a_j \left\langle \psi_n^0 \middle| H' \middle| \psi_j^0 \right\rangle \;+\; a_n \left\langle \psi_n^0 \middle| H' \middle| \psi_n^0 \right\rangle \;-\; E'_n a_n \;. \tag{10.26}$$

In Eq. (10.25), the sum is carried out for all values of j, whereas in Eq. (10.26), the sum is carried out for all values of j except the value $j = n$. Using the result of first-order perturbation theory for E_n' (see Eq. 10.17), then the last two terms of Eq. (10.26) cancel. Using the first-order perturbation result for a_j (Eq. 10.19), one obtains

$$\boxed{E''_n \;=\; \sum_{j\neq n} \frac{\left| \left\langle \psi_n^0 \middle| H' \middle| \psi_j^0 \right\rangle \right|^2}{E_n^0 \;-\; E_j^0}} \tag{10.27}$$

which is the second-order correction to the energy of the nth state of a system subjected to the perturbation hamiltonian H'. Note that the second-order term given in the above equation increases drastically if two energy-levels are closely spaced, *i. e.* for small $E_n{}^0 - E_j{}^0$. The second-order correction term becomes large for a small separation of two energy levels. It is therefore frequently said that energy levels *repel each other*. States energetically distant from the state of interest may be neglected in practical calculations.

For $m \neq n$ in Eq. (10.24), using that $a_n = 0$, and with Eqs. (10.17) and (10.19), one obtains

$m \neq n$:
$$b_m \;=\; \sum_{j\neq n} \frac{\left\langle \psi_j^0 \middle| H' \middle| \psi_n^0 \right\rangle \left\langle \psi_m^0 \middle| H' \middle| \psi_j^0 \right\rangle}{\left(E_n^0 - E_j^0\right)\left(E_n^0 - E_m^0\right)} \;-\; \frac{\left\langle \psi_n^0 \middle| H' \middle| \psi_n^0 \right\rangle \left\langle \psi_m^0 \middle| H' \middle| \psi_n^0 \right\rangle}{\left(E_n^0 - E_m^0\right)^2} \;. \tag{10.28}$$

Finally $b_m = b_n$ $(m = n)$ must be determined, which can again be achieved with the normalization condition.

$$m = n: \quad \left\langle \psi_n^0 + \lambda \sum_j a_j \psi_j^0 + \lambda^2 \sum_j b_j \psi_j^0 \middle| \psi_n^0 + \lambda \sum_j a_j \psi_j^0 + \lambda^2 \sum_j b_j \psi_j^0 \right\rangle = 1 \,. \qquad (10.29)$$

Evaluation of this integral yields the value of $b_{m=n} = b_n$

$$b_n = -\frac{1}{2} \sum_j |a_j|^2 = -\frac{1}{2} \sum_{j \neq n} \frac{\left| \left\langle \psi_j^0 \middle| H' \middle| \psi_n^0 \right\rangle \right|^2}{\left(E_n^0 - E_j^0 \right)^2} \,. \qquad (10.30)$$

Using these values of b_m, the second-order correction term of the wave function (see Eq. 10.22) is given by

$$\psi_n'' = \sum_{m \neq n} \left\{ \left[\sum_{j \neq n} \frac{\left\langle \psi_j^0 | H' | \psi_n^0 \right\rangle \left\langle \psi_m^0 | H' | \psi_j^0 \right\rangle}{\left(E_n^0 - E_j^0 \right)\left(E_n^0 - E_m^0 \right)} - \frac{\left\langle \psi_n^0 | H' | \psi_n^0 \right\rangle \left\langle \psi_m^0 | H' | \psi_n^0 \right\rangle}{\left(E_n^0 - E_m^0 \right)^2} \right] \psi_m^0 - \frac{\left| \left\langle \psi_m^0 | H' | \psi_n^0 \right\rangle \right|^2}{2 \left(E_n^0 - E_m^0 \right)^2} \psi_n^0 \right\} \qquad (10.31)$$

The second order correction terms for the energy and the wave function have now been obtained. For convenience, the results of first-order and of second-order perturbation theory for the energy and wave function of the *n*th state are summarized:

First- and second-order correction to the energy of the n*th *state:

$$E_n = E_n^0 + \lambda \left\langle \psi_n^0 \middle| H' \middle| \psi_n^0 \right\rangle + \lambda^2 \sum_{j \neq n} \frac{\left| \left\langle \psi_n^0 \middle| H' \middle| \psi_j^0 \right\rangle \right|^2}{E_n^0 - E_j^0} \qquad (10.32)$$

First- and second-order correction to the wave function of the n*th *state:

$$\psi_n = \psi_n^0 + \lambda \sum_{m \neq n} \frac{\left\langle \psi_m^0 \middle| H' \middle| \psi_n^0 \right\rangle}{E_n^0 - E_m^0} \psi_m^0 + \lambda^2 \sum_{m \neq n} \left\{ \left[\sum_{j \neq n} \frac{\left\langle \psi_j^0 | H' | \psi_n^0 \right\rangle \left\langle \psi_m^0 | H' | \psi_j^0 \right\rangle}{\left(E_n^0 - E_j^0 \right)\left(E_n^0 - E_m^0 \right)} - \frac{\left\langle \psi_n^0 | H' | \psi_n^0 \right\rangle \left\langle \psi_m^0 | H' | \psi_n^0 \right\rangle}{\left(E_n^0 - E_m^0 \right)^2} \right] \psi_m^0 - \frac{\left| \left\langle \psi_m^0 | H' | \psi_n^0 \right\rangle \right|^2}{2 \left(E_n^0 - E_m^0 \right)^2} \psi_n^0 \right\} \qquad (10.33)$$

10.3 Example for first-order perturbation calculation

A simple example for a first-order perturbation calculation is illustrated in ***Fig***. 10.2. The *unperturbed* potential is a square-shaped quantum well. The lowest eigenstate energy $E_{n=0}{}^0$ and wave function $\psi_{n=0}{}^0$ are shown as well. The system is now subjected to a potential perturbation, and it is the purpose of this example to calculate the perturbed ground-state energy

$E_{n=0} = E_{n=0}{}^0 + \lambda E_n'$. The perturbation used in this example is a potential energy perturbation given by

$$U_p(x) = 0 \qquad (x < b) \tag{10.34}$$

$$U_p(x) = -U_0 \qquad (b \le x \le c) \tag{10.35}$$

$$U_p(x) = 0 \qquad (x > c)\,. \tag{10.36}$$

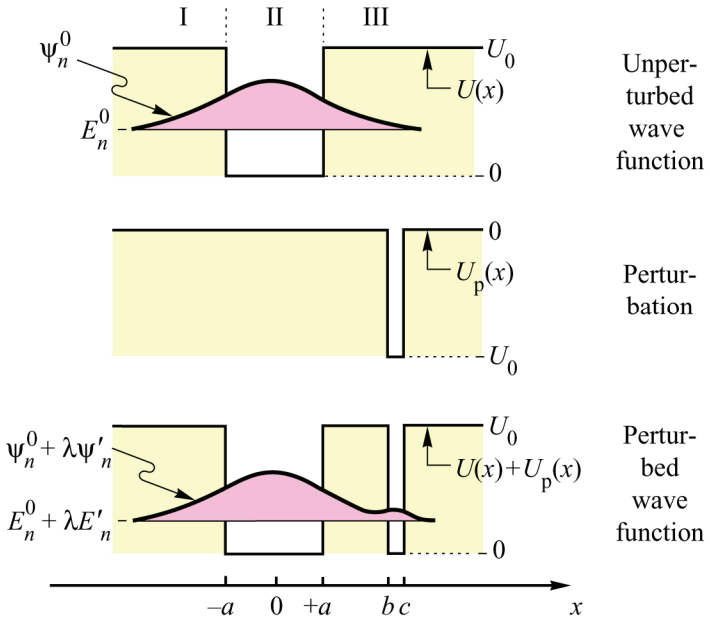


Fig. 10.2. Simple example of an (*i*) unperturbed wave function, (*ii*) a small perturbation, and (*iii*) and resulting perturbed wave function. The first-order correction to the energy due to the perturbation is $\lambda E'_n$. The correction to the wavefunction is $\lambda\psi'_n$.

Since the perturbation is purely a potential energy perturbation, the hamiltonian operator is given by

$$H = H^0 + H' = H^0 + U_p(x)\,. \tag{10.37}$$

The solution of the Schrödinger equation for the unperturbed problem (which is the finite square-well potential) is given by

$$\psi_I(x) = A_1\, e^{\kappa x} \tag{10.38}$$

$$\psi_{II}(x) = A_1 \left(\frac{e^{-\kappa a}}{\cos ka} \right) \cos kx \tag{10.39}$$

$$\psi_{III}(x) = A_1\, e^{-\kappa x} \tag{10.40}$$

where A_1 is a normalization constant and

$$\kappa = \frac{1}{\hbar}\sqrt{2m\left(U_0 - E_0^0\right)} \tag{10.41}$$

$$k = \frac{1}{\hbar}\sqrt{2m\,E_0^0} \tag{10.42}$$

where $E_{n=0}{}^0 = E_0{}^0$ is the lowest eigenstate energy of the unperturbed system.

The change of the lowest eigenstate energy is now calculated by using to Eq. (10.18). Furthermore, we use $\lambda = 1$ and obtain

$$E_0 = E_0^0 + \int_{-\infty}^{\infty} \psi_0^0\, U_\mathrm{p}(x)\, \psi_0^0 \; \mathrm{d}x\,. \tag{10.43}$$

Since $U_\mathrm{p}(x) = 0$ outside the interval between b and c, the integration can be limited to this interval

$$E_0 = E_0^0 + \int_b^c \left(-U_0\right) A_\mathrm{I}^2\, \mathrm{e}^{-2\kappa}\, \mathrm{d}x = E_0^0 - \frac{U_0\, A_\mathrm{I}^2}{2\,\kappa}\left(\mathrm{e}^{-2\kappa b} - \mathrm{e}^{-2\kappa c}\right). \tag{10.44}$$

The equation reveals that the perturbed energy of the state, E_0, is lower than the unperturbed value of the energy, $E_0{}^0$. Thus the eigenstate energy decreases upon perturbation.

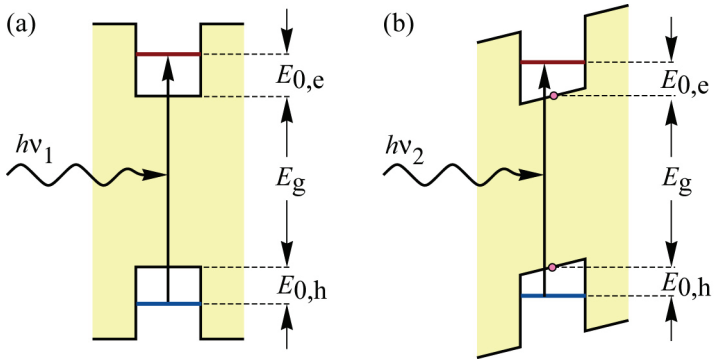


Fig. 10.3. Change in transition energy in a quantum well structure caused by an electric field. The transition energy sightly decreases with increasing electric field so that $h\nu_1 > h\nu_2$. (a) Structure without field. (b) Structure with field. The effect is known as the quantum-confined Sark effect.

Exercise 1: ***Quantum well electroabsorption*** **(*Quantum-confined Stark effect*)**. This exercise calculates the change in optical transition energy in a quantum well caused by an electric field. This effect is called the *quantum-confined Stark effect* and is applied in fast optical modulators. The change in transition energy is illustrated in ***Fig***. 10.3.

The detailed band diagram of a quantum well structure is shown in ***Fig***. 10.4. Consider an electron with effective mass m^* in a symmetric quantum well with thickness L_QW, clad by infinitely high barriers. Assume that the center of the quantum well is located at the origin of the coordinate system, *i. e.* at $x = 0$. The quantum well is now subjected to a constant electric field $\mathcal{E}$, so that the potential energy created by the electric field is given by $U(x) = e\,\mathcal{E}x$. Upon application of an electric field, the quantum well potential is perturbed. The perturbation hamiltonian for a constant electric field is given by

$$H' = -e\,\mathcal{E}\,x\,. \tag{10.45}$$

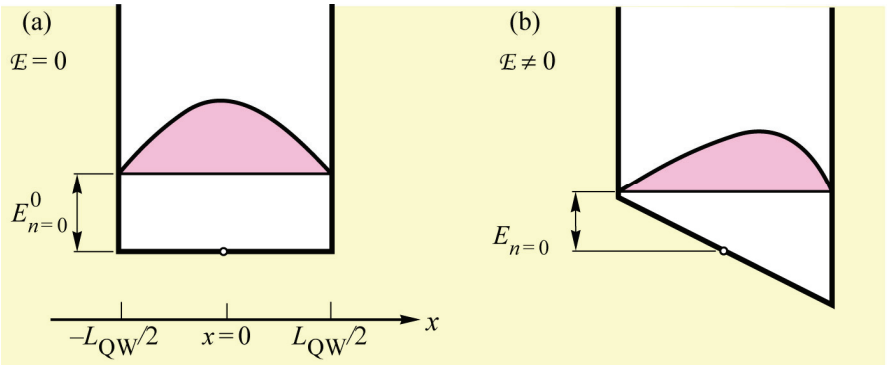


Fig. 10.4. Quantum well structure (a) without and (b) with electric field. The energy of the lowest state changes upon application of field. The change can be calculated by second-order perturbation theory.

(a) What is the energy of the lowest eigenstate in the well for $\mathcal{E} = 0$?
(b) Calculate the change in lowest eigenstate energy due to the electric field using first-order perturbation theory and show that first-order perturbation theory yields *no* change in the energy of the ground state.
(c) Calculate the change in lowest eigenstate energy due to the electric field using second-order perturbation theory. Use reasonable approximations and explain them. Show that second order perturbation theory yields a *decrease* in the ground-state energy.
(d) Show that the energy between the highest valence band state and the lowest conduction band state (*i. e*. the absorption edge) *decreases* upon application of an electric field.

Solution:

(a) With $\mathcal{E} = 0$, there is no perturbation and the wave functions are given by

$$\psi_n = \sqrt{\frac{2}{L_{QW}}} \cos\left(\frac{n+1}{L_{QW}}\pi x\right) \qquad n = 0, 2, 4 \ldots \qquad (10.46)$$

$$\psi_n = \sqrt{\frac{2}{L_{QW}}} \sin\left(\frac{n+1}{L_{QW}}\pi x\right) \qquad n = 1, 3, 5 \ldots \qquad (10.47)$$

The eigenstate energies are given by

$$E_0{}^0 = \frac{\hbar^2}{2m}\left(\frac{\pi}{L_{QW}}\right)^2 \qquad E_n = \frac{\hbar^2}{2m}\left[\frac{(n+1)\,\pi}{L_{QW}}\right]^2 \qquad n = 0, 1, 2, 3 \ldots \qquad (10.48)$$

(b) The first order correction due to the perturbation ($H' = -e\,\mathcal{E}\,x$) is given by

$$E_0{}' = \langle \psi_0 | H' | \psi_0 \rangle = \int_{-L/2}^{L/2} -e\,\mathcal{E}\,x \frac{2}{L_{QW}} \cos^2\left(\frac{\pi}{L_{QW}}x\right) dx = 0 \qquad (10.49)$$

Thus the first-order correction term is zero.

(c) According to Eq. (10.27), the second-order correction is given by

$$E_0'' = \sum_{n=1}^{\infty} \frac{\langle \psi_0 | H' | \psi_n \rangle}{E_0^0 - E_n^0} = \sum_{n=1}^{\infty} \frac{\langle \psi_0 | H' | \psi_{2n} \rangle}{E_0^0 - E_{2n}^0} + \sum_{n=1}^{\infty} \frac{\langle \psi_0 | H' | \psi_{2n-1} \rangle}{E_0^0 - E_{2n-1}^0} =$$

$$= 0 + \sum_{n=1}^{\infty} \frac{\langle \psi_0 | H' | \psi_{2n-1} \rangle}{\frac{\hbar^2}{2m}\left(\frac{\pi}{L_{QW}}\right)^2 - \frac{\hbar^2}{2m}\left(\frac{2n\pi}{L_{QW}}\right)^2} \qquad (10.50)$$

$$= \sum_{n=1}^{\infty} \frac{1}{\frac{\hbar^2}{2m}\left(\frac{\pi}{L_{QW}}\right)^2 \left(4n^2 - 1\right)} \frac{16 n e \mathcal{E} L_{QW} (-1)^n}{(2n+1)^2 (2n-1)^2} = \sum_{n=1}^{\infty} \frac{16 n e \mathcal{E} L_{QW} (-1)^n}{\frac{\hbar^2}{2m}\left(\frac{\pi}{L_{QW}}\right)^2 \left(4n^2 - 1\right)^3}$$

(d) Since the correction term is proportional to $(4n^2 - 1)^{-3}$, the *first* summand ($n = 1$) of the sum is *largest*. Thus the second-order correction term is dominated by the $n = 1$ term. Since the $n = 1$ summand is *negative*, second-order perturbation theory yields a *decrease* in ground state energy upon perturbation.

The concept of shifting the absorption edge by an electric field is used in electroabsorption quantum-well modulators. In these modulators, quantum wells are placed in the depletion region of a reverse-biased pn junction. Such electroabsorption modulators can be modulated at a much higher bit rate than would be possible by direct current-modulation of a semiconductor laser. The speed advantage is due to the much smaller depletion capacitance of the reverse-biased modulator junction as compared to the diffusion capacitance of the forward-biased laser junction.

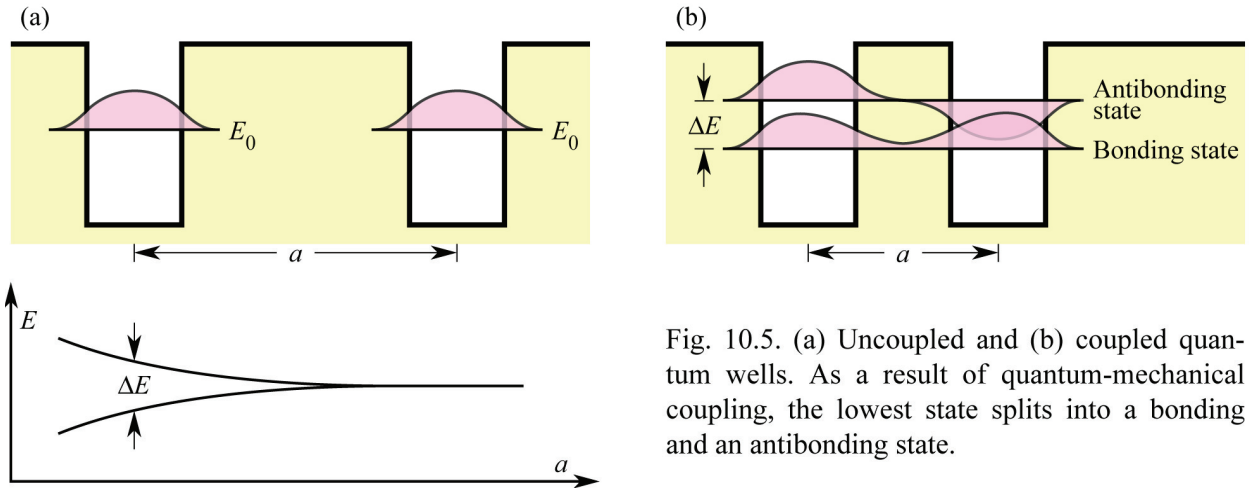


Fig. 10.5. (a) Uncoupled and (b) coupled quantum wells. As a result of quantum-mechanical coupling, the lowest state splits into a bonding and an antibonding state.

Exercise 2: ***Coupled quantum wells***. If the states of two identical quantum wells are perturbed by decreasing the distance between the quantum wells (QWs), the state will split into two states, a *bonding state* and an *anti-bonding* state. The energy split ΔE increases as the separation of the

QWs decreases. Coupled QWs and the schematic dependence of ΔE on the distance between the two QWs are shown in ***Fig***. 10.5. Give an estimate of the energy splitting.

Solution:
Using Eq. (10.44), the change in the quantum-state energy, due to the presence of an adjacent well, is given by

$$\Delta E = -\frac{U_0 A_1^2}{2\kappa}\left(e^{-2\kappa b} - e^{-2\kappa c}\right). \tag{10.51}$$

For a sufficiency thick well, $e^{-\kappa b} >> e^{-\kappa c}$, and the equation simplifies to

$$\Delta E = -\frac{U_0 A_1^2}{2\kappa} e^{-2\kappa b} \tag{10.52}$$

where we can identify the term $e^{-\kappa b}$ as the tunneling probability of an electron through the barrier. Assuming further that the energy splitting between the two coupled wells shown in ***Fig***. 10.5 is approximately equal to the perturbation energy calculated above, we can write

$$\Delta E \propto e^{-2\kappa b} \qquad \text{or} \qquad \Delta E \propto \text{tunneling probability} \tag{10.53}$$

Show the dependence of the energy levels of two identical atoms (e.g. hydrogen, H or oxygen, O) as a function of the distance between the two atoms. Explain the difference between the coupling of two QWs and the coupling of two atoms.

Solution:
When atoms get closer together, the valence electron orbitals start to overlap and the energy states split into two states which we call the ***bonding state*** and an ***anti-bonding state***. This is shown in ***Fig***. 10.6. At very small distances, the repulsive short-range potential of atoms causes the levels to increase in energy. The distance at which the bonding-state energy has the minimum has a special significance: This is the equilibrium distance between the two atoms of the molecule (e.g. H_2 or O_2).

The behavior of two coupled atoms is different from the two coupled quantum wells in that atoms have a repulsive short-range potential (while quantum wells do not).

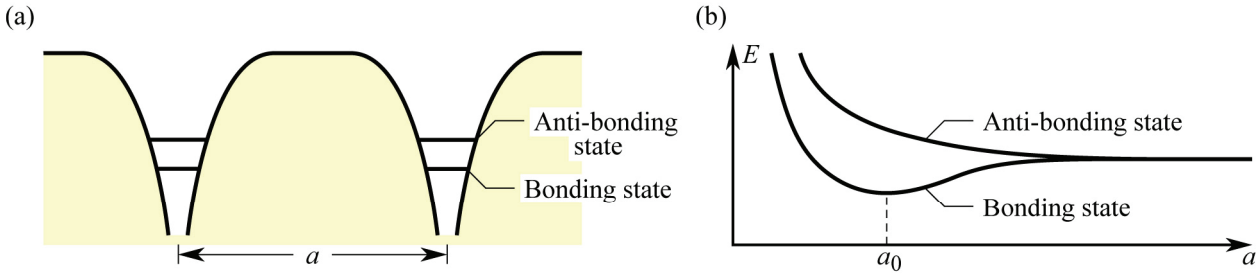


Fig. 10.6. (a) Potential energy and (b) energy levels of a molecule consisting of two identical atoms. The distance at which the bonding-state energy has a minimum, a_0, is the equilibrium distance between the atoms of the molecule.

11

Time-dependent perturbation theory

11.1 Time-dependent perturbation theory

In the preceding sections, we have considered time-independent quantum mechanical systems, in which wave functions and state energies do not depend on time. It is important to keep in mind that in such stationary systems, *nothing observable ever happens*. In the previous section it was calculated, how wave functions and the state energies change upon an external perturbation. Do such changes of wave functions occur instantaneously?

The answer to this question is no. A wave function $\Psi(x, t)$ represents a particle distribution of $\Psi^*(x, t)\ \Psi(x, t)$. To change the particle distribution, particles have to be moved (in a classical sense) which cannot be accomplished instantaneously. Therefore, the shape of a wave function cannot change instantaneously.

The role of time-dependent perturbation theory is further illustrated in ***Fig***. 11.1, where an infinite potential well is shown along with its ground state wave function. The perturbation, which is assumed to occur instantaneously at $t = t_0$, is a constant electric field superimposed to the infinite square-well potential. It is assumed that the potential energy in the center of the well remains constant before and after the perturbation as indicated in ***Fig***. 11.1. In the moment of the perturbation ($t = t_0$), the wave function continues to have its original shape. Not only the wave function remains the same, but also the expectation value of the *kinetic* energy remains unchanged for $t < t_0$ and $t = t_0$.

$$\Psi(x, t < t_0) \quad = \quad \Psi(x, t_0) \tag{11.1}$$

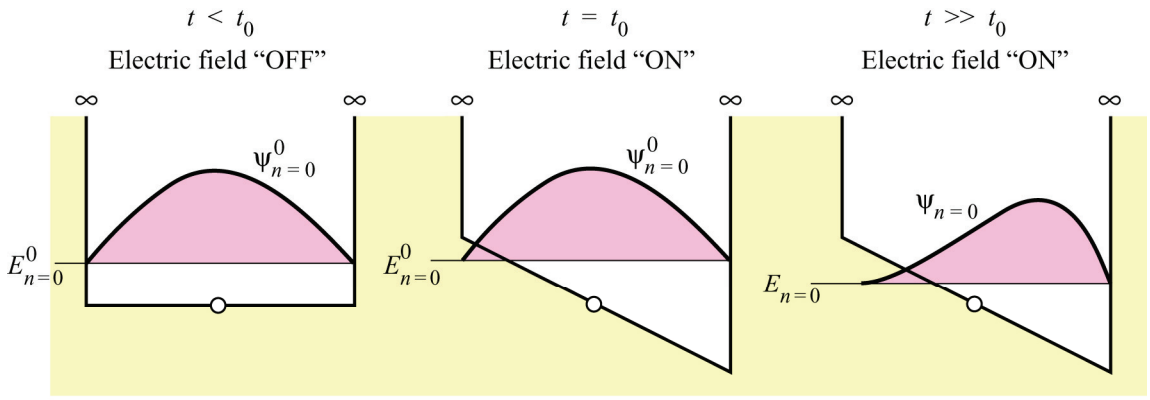


Fig. 11.1. Temporal evolution of a perturbation. The perturbation is assumed to occur instantly at the time t_0. The wavefunction at $t = t_0$ has still the shape as $\psi^0{}_{n=0}$ ($t < 0$). Then, the wavefunction changes according to the perturbation, and, it is given by $\psi_{n=0}$. The perturbation shown here is a constant electric field.

$$E_{\text{kin}}(t = t_0) \quad = \quad \left\langle \Psi(x, t_0) \left| \frac{p^2}{2m} \right| \Psi(x, t_0) \right\rangle . \tag{11.2}$$

The expectation value of the potential energy changes instantaneously if the potential perturbation occurs instantaneously. For the specific example shown in ***Fig***. 11.1, however, the potential energy of the particle

$$E_{\text{pot}}(t = t_0) \quad = \quad \left\langle \Psi(x, t_0) \left| U(x) \right| \Psi(x, t_0) \right\rangle \tag{11.3}$$

does not change since the potential perturbation is symmetric with respect to the center of the potential well.

It is the purpose of this section to provide an expression for the temporal evolution of the wave function when the system is exposed to a *time-dependent perturbation*. The starting point for obtaining the time-dependent one-dimensional wave function $\Psi = \Psi(x, t)$ is the time-dependent Schrödinger equation

$$H\,\Psi \quad = \quad -\frac{\hbar}{\mathrm{i}} \frac{\partial}{\partial t} \Psi \tag{11.4}$$

where H is the hamiltonian operator, which is composed of the time-independent part H^0 and the time-dependent part H'

$$H \quad = \quad H^0 \; + \; \lambda\, H'(t) . \tag{11.5}$$

The perturbation is assumed to be small, similar to the assumption made in time-independent perturbation theory. The time-dependent Schrödinger equation *before* the perturbation is given by

$$H^0\, \psi_n^0 \quad = \quad E_n^0\, \psi_n^0 . \tag{11.6}$$

Using the time-dependent wave function $\Psi_n{}^0 = \Psi_n{}^0 (x, t)$, Eq. (11.4) can be written as

$$H^0\, \Psi_n^0 \quad = \quad -\frac{\hbar}{\mathrm{i}} \frac{\partial}{\partial t} \Psi_n^0 \quad = \quad E_n^0\, \Psi_n^0 . \tag{11.7}$$

Here $\Psi_n{}^0$ is given by (see Chap. 6 for the derivation of the time dependence of $\Psi_n{}^0$)

$$\boxed{\Psi_n^0 \quad = \quad \psi_n^0\, \mathrm{e}^{-\mathrm{i}(E_n^0/\hbar)t}} \tag{11.8}$$

The general solution of the unperturbed system is given by the complete set of orthonormal solutions,

$$\Psi^0 \quad = \quad \sum\nolimits_n a_n\, \Psi_n^0 . \tag{11.9}$$

This general solution is normalized, if

$$\left\langle \Psi^0 \middle| \Psi^0 \right\rangle \;=\; 1 \qquad \text{or} \qquad \sum\nolimits_n a_n^* \, a_n \;=\; 1 \,. \tag{11.10}$$

To find the time-dependent wave function, the Schrödinger equation

$$\left[H^0 \;+\; \lambda \, H'(t) \right] \Psi \;=\; -\frac{\hbar}{\mathrm{i}} \frac{\partial}{\partial t} \Psi \tag{11.11}$$

must be solved. For this purpose, ***the perturbed wave function is expressed as the superposition of a complete set of orthonormal wave functions of the unperturbed system***

$$\Psi(x,t) \;=\; \sum\nolimits_n a_n(t) \, \Psi_n^0(x,t) \,. \tag{11.12}$$

Because $\Psi(x, t)$ is represented by a *complete* set of orthonormal wave functions, any wave function can be represented by this set. It may seem that Eq. (11.12) is rather complicated, since both, $a_n(t)$ and $\Psi_n^0(x, t)$, depend on the variable *time*. However, the method used here, *i. e.* the method of *variation of parameters* (also called *variation of constants*) is quite powerful and provides the general solution to the problem.

Substitution of Eq. (11.12) into the perturbed wave equation (Eq. 11.11) yields

$$\sum_n a_n(t) \, H^0 \, \Psi_n^0 + \sum_n a_n(t) \, \lambda \, H' \, \Psi_n^0 \;=\; -\frac{\hbar}{\mathrm{i}} \sum_n \left[\frac{\partial}{\partial t} \, a_n(t) \right] \Psi_n^0 - \frac{\hbar}{\mathrm{i}} \sum_n a_n(t) \, \frac{\partial}{\partial t} \Psi_n^0 \,. \tag{11.13}$$

The first summand of the left-hand side and the second summand of the right-hand side of this equation are equal; these two terms represent the unperturbed Schrödinger equation and they can be removed from the equation. Now consider one specific wave function Ψ_m^0 of the orthonormal set Ψ_n^0. Pre-multiplication of the remainder of the equation with Ψ_m^{0*} followed by integration (recall that $\langle \Psi_m^0 \,|\, \Psi_n^0 \rangle = \delta_{mn}$) yields

$$\boxed{ \frac{\mathrm{d}}{\mathrm{d}t} \, a_m(t) \;=\; -\frac{\mathrm{i}}{\hbar} \lambda \sum\nolimits_n a_n(t) \left\langle \Psi_m^0 \middle| H' \middle| \Psi_n^0 \right\rangle } \tag{11.14}$$

This is the ***fundamental result of time-dependent perturbation theory***. The result gives the rate of change of the *m*th parameter, a_m. Knowing all parameters a_n, allows us to describe the actual wave function of the perturbed system (see Eq. 11.12).

Let us have a closer look at the fundamental result of time-dependent perturbation theory (Eq. 11.14). The rate of change of a_m depends upon the amplitudes of all the other a_n, and it also depends on a set of ***matrix elements*** $\langle \Psi_m^0 | \, H' \, | \, \Psi_n^0 \rangle$. Since the sum adds over all states with subscript *n*, a *set of matrix elements connects* the state Ψ_m^0, via the perturbation hamiltonian H', with all other wave functions Ψ_n^0.

Although, Eq. (11.14) may look quite simple, it stands for a large system of equations. For illustration, a part of the system of linear equation is written out:

$$
\begin{aligned}
-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}a_0}{\mathrm{d}t} &= a_0\left\langle\Psi_0^0\middle|H'\middle|\Psi_0^0\right\rangle + a_1\left\langle\Psi_0^0\middle|H'\middle|\Psi_1^0\right\rangle + \cdots + a_j\left\langle\Psi_0^0\middle|H'\middle|\Psi_j^0\right\rangle + \cdots \\
-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}a_1}{\mathrm{d}t} &= a_0\left\langle\Psi_1^0\middle|H'\middle|\Psi_0^0\right\rangle + a_1\left\langle\Psi_1^0\middle|H'\middle|\Psi_1^0\right\rangle + \cdots + a_j\left\langle\Psi_1^0\middle|H'\middle|\Psi_j^0\right\rangle + \cdots \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad (11.15) \\
-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}a_j}{\mathrm{d}t} &= a_0\left\langle\Psi_j^0\middle|H'\middle|\Psi_0^0\right\rangle + a_1\left\langle\Psi_j^0\middle|H'\middle|\Psi_1^0\right\rangle + \cdots + a_j\left\langle\Psi_j^0\middle|H'\middle|\Psi_j^0\right\rangle + \cdots \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots
\end{aligned}
$$

Both equations, Eq. (11.14) and Eq. (11.15) are identical and therefore have the same physical content. However, the more explicit form of Eq. (11.15) is better suited to appreciate and understand the fundamental result of time-dependent perturbation theory. In general, if a physical system has a large number of eigenstates, than Eq. (11.15) consists of a large number of equations.

It is obviously very tedious to solve the system of equations given in Eqs. (11.15). However, these equations provide the mathematically *accurate* solution of the time-dependent wave function. To simplify the solution of the system of equations, a first-order approximation for $a_j(t)$ will be used; to do so, $a_j(t)$ is expanded into a power series of λ

$$\boxed{a_j = a_j^0 + \lambda a_j' + \lambda^2 a_j'' + \cdots} \qquad (11.16)$$

where $a_j' = \mathrm{d}a_j/\mathrm{d}\lambda$ and $a_j'' = \mathrm{d}^2a_j/\mathrm{d}\lambda^2$. This is the same expansion that was used in time-independent perturbation theory as discussed in the preceding chapter. The *linear* term in Eq. (11.16) is assumed to be the largest correction, while the squared and all higher terms are assumed to be vanishingly small. The *linear* term ($\lambda a_j'$) in Eq. (11.16) is the term of interest for first-order time-dependent perturbation theory. Therefore, the square term ($\lambda^2 a_j''$) and all higher terms will be neglected. Substitution of Eq. (11.16) into Eq. (11.15) yields for a_j

$$
\begin{aligned}
-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}}{\mathrm{d}t}\left(a_j^0 + \lambda a_j' + \lambda^2 a_j''\right) &= \left(a_0^0 + \lambda a_0' + \lambda^2 a_0''\right)\left\langle\Psi_j^0\middle|\lambda H'\middle|\Psi_0^0\right\rangle \\
&\quad + \cdots \qquad (11.17) \\
&\quad + \left(a_j^0 + \lambda a_j' + \lambda^2 a_j''\right)\left\langle\Psi_j^0\middle|\lambda H'\middle|\Psi_j^0\right\rangle \\
&\quad + \cdots
\end{aligned}
$$

If $\lambda = 0$ (no perturbation) then all the summands on the right-hand side of the equation are zero. Therefore

$$\frac{\mathrm{d}}{\mathrm{d}t}a_0^0 = 0; \qquad \frac{\mathrm{d}}{\mathrm{d}t}a_1^0 = 0; \qquad \frac{\mathrm{d}}{\mathrm{d}t}a_2^0 = 0; \qquad \cdots \qquad (11.18)$$

That is, in the absence of a perturbation ($\lambda = 0$), all a_j^0's are stationary, that is, they do not

depend on time. In other words, in the absence of a time-dependent perturbation, the wave functions are steady-state functions.

We next consider only the *first-order* correction terms and will *neglect* the *second-order* correction terms. Equating all terms for λ^1 in Eq. (11.17) yields the following system of equations.

$$\begin{aligned}
-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}a_0'}{\mathrm{d}t} &= a_0^0\left\langle\Psi_0^0\middle|H'\middle|\Psi_0^0\right\rangle + a_1^0\left\langle\Psi_0^0\middle|H'\middle|\Psi_1^0\right\rangle + \cdots + a_j^0\left\langle\Psi_0^0\middle|H'\middle|\Psi_j^0\right\rangle + \cdots \\
-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}a_1'}{\mathrm{d}t} &= a_0^0\left\langle\Psi_1^0\middle|H'\middle|\Psi_0^0\right\rangle + a_1^0\left\langle\Psi_1^0\middle|H'\middle|\Psi_1^0\right\rangle + \cdots + a_j^0\left\langle\Psi_1^0\middle|H'\middle|\Psi_j^0\right\rangle + \cdots \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad (11.19) \\
-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}a_j'}{\mathrm{d}t} &= a_0^0\left\langle\Psi_j^0\middle|H'\middle|\Psi_0^0\right\rangle + a_1^0\left\langle\Psi_j^0\middle|H'\middle|\Psi_1^0\right\rangle + \cdots + a_j^0\left\langle\Psi_j^0\middle|H'\middle|\Psi_j^0\right\rangle + \cdots \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots
\end{aligned}$$

This set of equation represents an *approximate* first-order solution of the perturbed problem. The great advantage of this set of equations is that the perturbed wave function can be calculated from the stationary, unperturbed parameters, a_j^0. The a_j^0 thus represent the initial conditions of the problem. Equation (11.19) allows us to determine the growth and decline of each individual eigenfunction of the system.

In an effort to further understand the result of Eq. (11.19), we consider the wave function Ψ_0^0. This wave function is "connected" via H' to all other wave functions, Ψ_j^0. As time proceeds, the state Ψ_j^0 *feeds amplitude* into the state Ψ_0^0 at a rate given by $a_j^0\langle\Psi_0^0|H'|\Psi_j^0\rangle$; the reverse process *feeds amplitude* at rate $a_0^0\langle\Psi_j^0|H'|\Psi_0^0\rangle$. Thus, the perturbation H' constantly redistributes the amplitudes between all eigenfunctions.

What is the physical meaning of a quantum system whose state amplitudes vary with time? This question is easily answered by recalling that $\Psi^*\Psi$ is the probability amplitude of the spatial distribution of a quantum mechanical particle. Decreasing the amplitude of one wave function and simultaneously increasing the amplitude of the wave function of another state simply means that the quantum particle, which prior to the perturbation occupied one state, is transferred to another state as a result of the perturbation.

11.2 Step-function-like perturbation

Time-dependent perturbation theory is further simplified, if the initial state is represented by a single wave function Ψ_j^0. ***That is, the quantum-mechanical particle initially occupies only one single state, we assume here the jth state***. The *j*th coefficient in Eq. (11.12) then has a value of $a_j = 1$ and all other coefficients $a_{m\neq j} = 0$. Furthermore only the *j*th column of Eq. (11.19) is non-zero, while all other columns are zero. Let us further assume that the perturbation is *off* for $t < 0$ and *on* for $t \geq 0$ but would have no further time dependence. That is, the perturbation can be described by

$$H'(t) = H'\,\sigma(t) \qquad (11.20)$$

where $\sigma(t)$ is the step function. This function can assume values of either 0 or 1. It is $\sigma(t < 0) = 0$ and $\sigma(t \geq 0) = 1$ as shown in ***Fig***. 11.2. The dependence of H' on the spatial coordinates, *i. e.* $H' = H'(x)$, is unchanged.

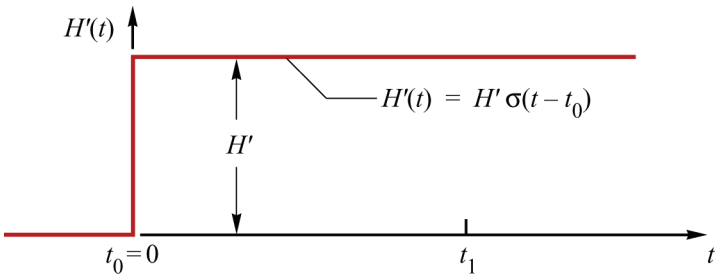


Fig. 11.2. Time dependence of step-function-like perturbation $H'(t)$. It is $H'(t \geq t_0) = H' =$ constant.

With these conditions, only the jth column of Eq. (11.19) is non-zero. With $\Psi_n^{\,0} = \psi_n^{\,0} \exp(-\mathrm{i}\, E_n\, t / \hbar)$, the jth column becomes

$$
\begin{aligned}
-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}}{\mathrm{d}t}\, a_0' &= \left\langle \psi_0^0 \middle| H' \middle| \psi_j^0 \right\rangle \mathrm{e}^{\mathrm{i}\omega_{0j}t} \\
-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}}{\mathrm{d}t}\, a_1' &= \left\langle \psi_1^0 \middle| H' \middle| \psi_j^0 \right\rangle \mathrm{e}^{\mathrm{i}\omega_{1j}t} \qquad (11.21) \\
a_j' &\approx 0 \\
-\frac{\hbar}{\mathrm{i}}\frac{\mathrm{d}}{\mathrm{d}t}\, a_m' &= \left\langle \psi_m^0 \middle| H' \middle| \psi_j^0 \right\rangle \mathrm{e}^{\mathrm{i}\omega_{mj}t}
\end{aligned}
$$

where $\omega_{mj} = (E_m^{\,0} - E_j^{\,0}) / \hbar$. Note that a_j' must be approximately zero, since $a_j = 1$ at $t < 0$ due to the assumption made above, and it is $a_j \approx 1$ for $t \geq 0$ due to the smallness of the perturbation. Therefore it is $a_j' \approx 0$ for all times considered here. If $a_m' = 0$ for $t < 0$, then all other equations of the system Eq. (11.21) are formally identical and have the same solution. This solution is obtained by integration over t. The integration constant of the integration is chosen in such a way to obtain $a_m'(t = 0) = 0$, which guarantees that the system is, at $t = 0$, still in its unperturbed state. The integration then yields

$$
a_m'(t_1) = -\frac{1}{\hbar}\left\langle \psi_m^0 \middle| H' \middle| \psi_j^0 \right\rangle \frac{\mathrm{e}^{\mathrm{i}\omega_{mj}t_1} - 1}{\omega_{mj}} \qquad (m = 0,\, 1,\, 2 \ldots\, m \neq j \ldots)\,. \qquad (11.22)
$$

Thus, at a time $t = t_1$, the amplitudes of all states are non-zero, $a_m'(t_1) \neq 0$, even though $a_m'(t_1 \leq 0) = 0$. Note that, due to the smallness of the perturbation, $a_{m=j}' \approx 1$, and $a_{m\neq j}' \approx 0$ for all times. The result obtained thus provides the time development of the coefficients a_m, which determine the temporal development of the wave functions of the system. Note that the $a_m'(t_1)$ are calculated from *known* wave functions of the *unperturbed* problem plus the *known* perturbation hamiltonian.

If expectation values are calculated, then the term $a_m'^{*} a_m'$ rather than a_m' will become important. Modification of Eq. (11.22) by means of trigonometric conversions yields

$$a_m'^{*}\, a_m' \;=\; \frac{\left\langle \psi_m^0 \middle| H' \middle| \psi_j^0 \right\rangle^{*} \left\langle \psi_m^0 \middle| H' \middle| \psi_j^0 \right\rangle \sin^2\left(\frac{1}{2}\,\omega_{mj}\; t_1\right)}{\hbar^2 \left(\frac{1}{2}\,\omega_{mj}\right)^2} \qquad (m = 0,\, 1,\, 2 \;...\; m \neq j \;...)\,. \tag{11.23}$$

This equation gives the temporal dependence of the intensity, *i. e*. squared amplitude, of a wave function. As mentioned earlier, this equation is valid, if only the *j*th state wave function is non-zero before the occurrence of the perturbation.

We will now examine the results qualitatively to obtain a better understanding of time-dependent perturbations. Let's again assume to have a step-function-like perturbation. Initially, only the *j*th state is occupied and all other states are assumed to be empty. After the perturbation is switched on, the amplitudes of wave functions corresponding to other states increase, as inferred from Eq. (11.22). However, the a_m will increase only if $\langle \Psi_m^{\,0} | H' | \Psi_j^{\,0} \rangle \neq 0$. Let us assume that $\langle \Psi_m^{\,0} | H' | \Psi_j^{\,0} \rangle$ is in fact non-zero, then $a_m(t)$ increases with time, after the perturbation is switched on. In other words, the perturbation causes a redistribution of the amplitudes of the wave functions. In a particle-oriented point of view, one would state, that a particle originally occupying the *j*th state of energy E_j has some probability to transfer to the *m*th state with a different total energy E_m.

It is helpful to draw a comparison with classical mechanics. Assume a body moving frictionless on a plane at a constant velocity. Upon a "perturbation", for example a dip or hill in the potential plane, the velocity of the body will change, *i. e*., its kinetic energy. We now include the quantum-mechanics into this classical picture, namely (*i*) the *discreteness* of quantum energies and (*ii*) the uncertainty principle. Inclusion of these two principles brings about the quantum mechanical picture: Instead of a continuous change of the dynamical variable *energy* in the classical picture, we obtain in the quantum mechanical picture a *redistribution of probabilities between states with discrete energies*.

A perturbation can result in a *very selective redistribution* of probabilities, that is, the wave functions of some states may remain completely unaffected. To illustrate this, we consider the quantity $\langle \Psi_m^{\,0} | H' | \Psi_j^{\,0} \rangle$, in which the wave function $\Psi_j^{\,0}$ (of the initially populated *j*th state) is connected via the perturbation hamiltonian H' to the wave function $\Psi_m^{\,0}$ (of the initially unpopulated state). The quantity

$$\boxed{H'_{mj} \;=\; \left\langle \psi_m^0 \middle| H' \middle| \psi_j^0 \right\rangle} \tag{11.24}$$

is called the ***transition matrix element*** of the two wave functions $\Psi_m^{\,0}$ and $\Psi_j^{\,0}$. This matrix element may be either zero or non-zero. Assume, for example, that $\Psi_m^{\,0}$ and H' depend on the spatial coordinate *x* and that they are *even* functions with respect to *x*. Assume further that $\Psi_j^{\,0}$ is an odd function with respect to *x*. Then the matrix element is given by

$$H'_{mj} \;=\; \int_{-\infty}^{\infty} \psi_m^{0*}(x)\, H'(x)\, \psi_j^0(x)\, \mathrm{d}x\,. \tag{11.25}$$

The integrand will be an odd function and the integration yields $H_{mj}' = 0$. Thus, a perturbation does not affect the wave function of the *m*th state at all, if the corresponding matrix element H_{mj}' is zero. The matrix element H_{mj}' is the source of the ***selection rules*** of atomic, nuclear or semiconductor quantum well spectra. In such spectra some transitions are *allowed* ($H_{mj}' \neq 0$)

while other transitions are *forbidden* or *disallowed* ($H_{mj}' = 0$).

11.3 Harmonic perturbation and Fermi's Golden Rule

We next consider the case of a harmonically periodic perturbation. The case of a harmonic perturbation is of great importance, since many excitations, *e. g*. electromagnetic waves, are harmonic. We assume that the perturbation is step-function-like, that is, it is *switched on* at $t = 0$. The perturbation is then given by

$$H' = 0 \qquad (t < 0) \tag{11.26}$$

$$H' = A(x)\left(e^{i\omega_0 t} + e^{-i\omega_0 t}\right) \qquad (t \geq 0)\,. \tag{11.27}$$

In principle, any sinusoidal perturbation can be assumed in this equation. However, a perturbation of the form $(e^{i\omega_0 t} + e^{-i\omega_0 t})$ facilitates the following calculation. Furthermore, we have assumed that the perturbation hamiltonian depends only on the spatial coordinate x. This dependence is given by the function $A(x)$. A *harmonic* step-function-like perturbation rather than a *constant* step-function-like perturbation further simplifies Eq. (11.22). Insertion of Eq. (11.27) into the mth equation of Eq. (11.21) yields

$$-\frac{\hbar}{i}\frac{d}{dt}a_m' = H_{mj}'\, e^{i\omega_{mj}t}\left(e^{i\omega_0 t} + e^{-i\omega_0 t}\right) \qquad (m = 0, 1, 2 \ldots m \neq j \ldots) \tag{11.28}$$

where H_{mj}' is the previously defined matrix element

$$H_{mj}' = \left\langle \psi_m^0 \middle| H' \middle| \psi_j^0 \right\rangle = \int_{-\infty}^{\infty} \psi_m^{0*}(x)\; A(x)\; \psi_j^0(x)\; dx\,. \tag{11.29}$$

If the mth state is unoccupied at $t = 0$, *i. e*., $a_m'(t = 0) = 0$. Then $(d/dt)\, a_m'$ in Eq. (11.28) can be integrated with respect to t and one obtains at a time t_1 after the perturbation was switched on

$$a_m'(t_1) = -\frac{H_{mj}'}{\hbar}\left[\frac{\left(e^{i(\omega_{mj}+\omega_0)t_1} - 1\right)}{\omega_{mj} + \omega_0} - \frac{\left(e^{i(\omega_{mj}-\omega_0)t_1} - 1\right)}{\omega_{mj} - \omega_0}\right]. \tag{11.30}$$

Inspection of this equation indicates that $a_m'(t_1)$ becomes very large, if one of the denominators vanishes. That is, the mth state becomes strongly excited, if $\omega_{mj} = \pm\,\omega_0$, that is, if the energy associated with the perturbation ($\hbar\omega_0$) coincides with the energy difference between the mth state and the initially populated jth state ($\hbar\omega_m - \hbar\omega_j$). Thus, the states most strongly affected by the perturbation are those at energy $E_m^{\,0} = E_j^{\,0} \pm \hbar\omega_0$. The states between and beyond these two energies are excited as well, however with a weaker intensity. In the present simplified picture one obtains $a_m' \to \infty$ as $\hbar\omega_{mj} \to \pm\hbar\omega_0$. In reality, a_m' assumes large but finite values. Note that a_m' is limited to values << 1, since we have assumed a *small* perturbation.

To determine the magnitude of excitation of the mth state at a time t after the perturbation was switched on, we must calculate $a_m'^*(t)\, a_m'(t)$. (Note that we used the time t_1 in Eq. 11.30 to emphasize that t_1 is a specific time after the perturbation was switched on. We now re-name the variable t_1 and just use t to denote a specific time after the perturbation has been switched on.)

The calculation of this product provides two approximate expressions in the vicinity of the resonance energies. We obtain for $E_m{}^0 \approx E_j{}^0 + \hbar\omega_0$ (we call such transitions ***excitation transitions***, or, if the stimulus is optical, an ***absorption transition***):

$$a'^*_m(t)\, a'_m(t) \quad \approx \quad \frac{4\, H'^*_{mj}\, H'_{mj}}{\hbar^2}\, \frac{\sin^2\left[\frac{1}{2}\left(\omega_{mj} - \omega_0\right)t\right]}{\left(\omega_{mj} - \omega_0\right)^2} \,. \tag{11.31}$$

And in the vicinity of $E_m{}^0 \approx E_j{}^0 - \hbar\omega_0$ (we call such transitions ***de-excitation transitions***, or, if the stimulus is optical, a ***stimulated-emission transition***):

$$a'^*_m(t)\, a'_m(t) \quad \approx \quad \frac{4\, H'^*_{mj}\, H'_{mj}}{\hbar^2}\, \frac{\sin^2\left[\frac{1}{2}\left(\omega_{mj} + \omega_0\right)t\right]}{\left(\omega_{mj} + \omega_0\right)^2} \,. \tag{11.32}$$

The temporal development of the intensity of the wave function, $a_m'^*(t)\, a_m'(t)$, is shown in ***Fig***. 11.3. As illustrated in this figure, the system is initially characterized by a single state of energy E_j. After harmonic excitation, the intensity of wave functions of energy $E_j \pm \hbar\omega_0$ increases steadily. For the sake of simplicity, the matrix element was assumed to equal unity. Note that the matrix element H_{mj}' has a significant *selective effect* on the excitation of individual states. Initially, some states may not become excited at all. However, if the perturbation persists *very long*, then states may become excited via intermediate states. That is, a state in which assumed to have a zero matrix element with state j, $H_{mj}' = 0$ may become excited via an intermediate state k, if $H_{kj}' \neq 0$ and $H_{mk}' \neq 0$.

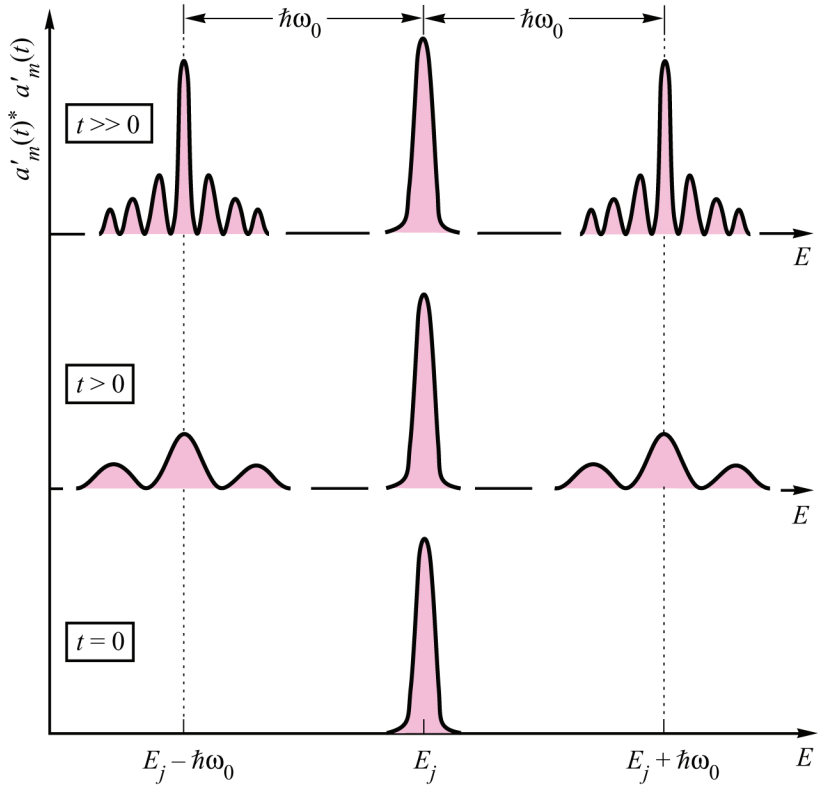


Fig. 11.3. Temporal development of intensities of wave functions. Initially the system is represented by a single occupied state with energy E_j (t = 0). After a harmonic perturbation with energy $(h/2\pi)\omega_0$, states of energy $E_j \pm(h/2\pi)\omega_0$ become increasingly populated. A transition to the state with energy $E_j + (h/2\pi)\omega_0$ is a *stimulated upward* (excitation) transition or *stimulated absorption* transition. Correspondingly, a quantum mechanical transition to the state with energy $E_j - (h/2\pi)\omega_0$ is a *stimulated downward* (de-excitation) transition or *stimulated emission* transition.

Frequently, a transition occurs between a single state and a *group of states* clustered about a state m with $E_m^{\,0} > E_j^{\,0}$. Transitions between an impurity state and band states in semiconductors are an example of such a transition. Let the cluster of states be characterized by a density of states $\rho(\omega_{mj})$ per unit of ω_{mj}. Assuming that the transition is an *excitation transition* in which the final state energy is higher than the initial state energy, *i. e.*, $E_m^{\,0} \approx E_j^{\,0} + \hbar\omega_0$, then Eq. (11.31) becomes

$$\left|a'_m(t)\right|^2 \;=\; \frac{1}{\hbar^2}\int_{-\infty}^{\infty}\left|H'_{mj}\right|^2 \frac{\sin^2\left[\frac{1}{2}\left(\omega_{mj}\;-\;\omega_0\right)t\right]}{\left[\frac{1}{2}\left(\omega_{mj}\;-\;\omega_0\right)\right]^2}\;\rho(\omega_{mj})\;\mathrm{d}\omega_{mj}\;. \qquad (11.33)$$

If $|H_{mj}'|^2$ is not a strong function of the final state m, then we can take it outside of the integral. The remaining integral is a product of two functions, namely the density of states

$$\rho(\omega_{mj}) \qquad (11.34)$$

and

$$\frac{\sin^2\left[\frac{1}{2}\left(\omega_{mj}\;-\;\omega_0\right)t\right]}{\left[\frac{1}{2}\left(\omega_{mj}\;-\;\omega_0\right)\right]^2}\;. \qquad (11.35)$$

These two functions are shown in ***Fig***. 11.4. The width of the latter function is approximately $2\pi/t$, and the width can be made arbitrarily small by increasing the observation time. As the width of the function decreases with time, we consider a time sufficiently long so that the width $2\pi/t$ is much smaller than that of the density of state function $\rho(\omega_{mj})$. This situation is shown in ***Fig***. 11.4. The area under the curve can be calculated by the integral

$$\int_{-\infty}^{\infty}\frac{\sin^2\left[\frac{1}{2}\left(\omega_{mj}\;-\;\omega_0\right)t\right]}{\left[\frac{1}{2}\left(\omega_{mj}\;-\;\omega_0\right)\right]^2}\,\mathrm{d}\omega_{mj} \;=\; 2\pi t\;. \qquad (11.36)$$

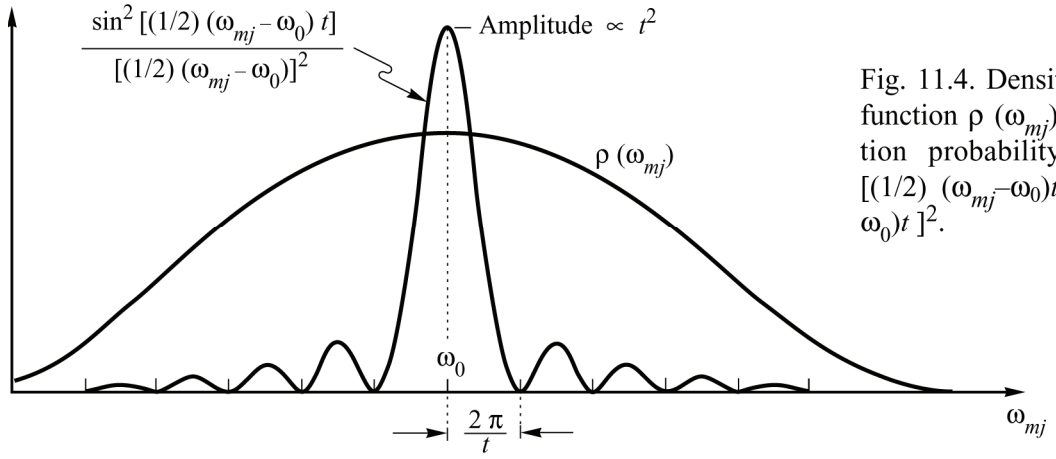


Fig. 11.4. Density of final states function $\rho(\omega_{mj})$ and the transition probability function $\sin^2[(1/2)(\omega_{mj}-\omega_0)t]/[(1/2)(\omega_{mj}-\omega_0)t]^2$.

Hence, the integral in Eq. (11.33) becomes

$$\left|a'_m(t)\right|^2 \;=\; \frac{2\pi}{\hbar^2}\,\left|H'_{mj}\right|^2\;\rho(\omega_{mj}=\omega_0)\;t\;. \tag{11.37}$$

Recalling that $\psi^*\,\psi$ is the probability density, $|\,a_m'\,|^2$ is the probability that the system is in the state m. Accordingly, the transition probability from a state j to a state m can be calculated from the change of $|\,a_j'\,|^2$ and $|\,a_m'\,|^2$ with time. The ***transition probability*** of a system from the state j to the state m is given by

Fermi's Golden Rule

$$\boxed{W_{j\to m} \;=\; \frac{\mathrm{d}}{\mathrm{d}t}\left|a'_m(t)\right|^2 \;=\; \frac{2\pi}{\hbar}\left|H'_{mj}\right|^2\;\rho(E=E_j+\hbar\,\omega_0)} \tag{11.38}$$

where $\rho(E)$ is the density of final states expressed as a function of energy using $\rho(\omega)=\hbar\,\rho(E)$. Enrico Fermi called Eq. (11.38) the ***Golden Rule*** of time-dependent perturbation theory because it plays a fundamental part in many applications. Fermi's Golden Rule was derived for an *excitation* transition, *i. e.* $E_m > E_j$. For *de-excitation* transitions, *i. e.* $E_m < E_j$, the density of states in Eq. (11.82), $\rho(E=E_j+\hbar\,\omega_0)$, must be replaced by $\rho(E=E_j-\hbar\,\omega_0)$.

We next consider a transition from one state to another single state within a continuum of states. In this case, Eq. (11.38) can be written as

$$W_{j\to m} \;=\; \frac{\mathrm{d}}{\mathrm{d}t}\left|\,a'_m(t)\,\right|^2 \;=\; \frac{2\pi}{\hbar}\left|H'_{mj}\right|^2\,\delta\!\left(E_m-E_j-\hbar\,\omega_0\,\right). \tag{11.39}$$

The state density is reduced to a single state, *i. e.* the δ function. Eq. (11.39) can be formally obtained from Eq. (11.33) by using $\sin^2(\xi\,t/2)/(\xi/2)^2 \to 2\pi\,t\,\delta(\xi)$ which is valid for sufficiently long t.

Two assumptions were made in deriving Fermi's Golden Rule. The first was that $2\pi/t$ be small compared with the width of the density of states function $\rho(\omega_{mj})$ as shown in ***Fig***. 11.4. The second assumption was $|\,a_m'(t)\,|^2 << 1$ which allows us to use of first-order perturbation theory. Otherwise second and higher order terms must be taken into account. Using either Eq. (11.31) or (11.32), the second condition can be stated as

$$\frac{\left|H'_{mj}\right|}{\hbar} \;<<\; \frac{1}{t} \tag{11.40}$$

where we have used that $\sin(\xi\,t)\approx\xi\,t$ which is valid for sufficiently small times. The physical significance is that the results of first-order perturbation theory are only valid for times short enough so that the transition probability out of the initial state j is small compared with unity.

Exercise 1: ***Band-to-band transitions in semiconductors***. In semiconductors, the valence band wave functions derive themselves from atomic orbitals with p-like symmetry and valence band wave functions are, therefore, of odd symmetry with respect to the position of the atom core. On the other hand, conduction band wave functions derive themselves from atomic orbitals with s-like symmetry and conduction band wave functions are, therefore, even functions with respect to the atom core. ***Fig***. 11.5 shows the atomic-scale bulk and quantum well wave functions in a

semiconductor.

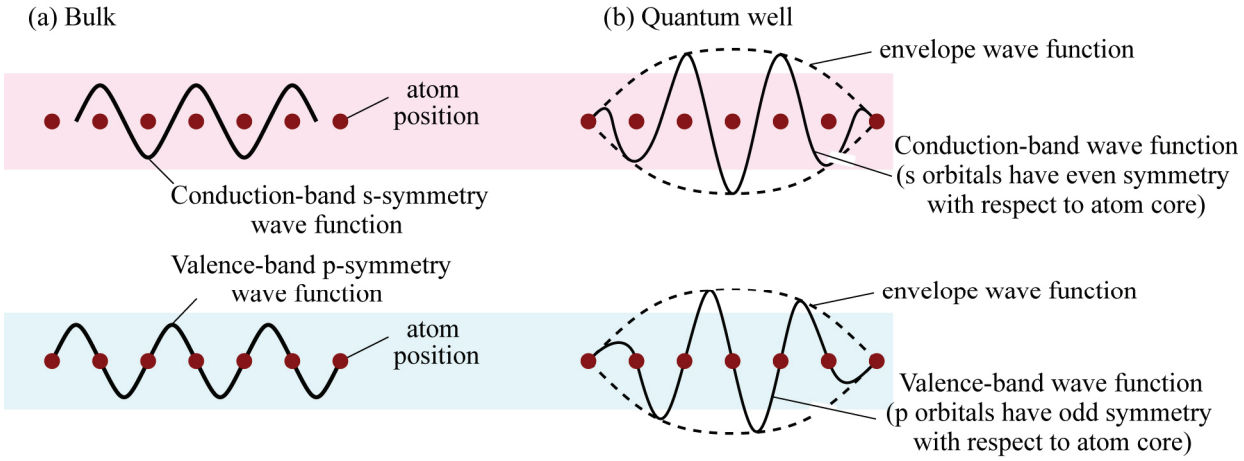


Fig. 11.5. Conduction and valence band wave functions of (a) bulk semiconductors and (b) quantum well structures.

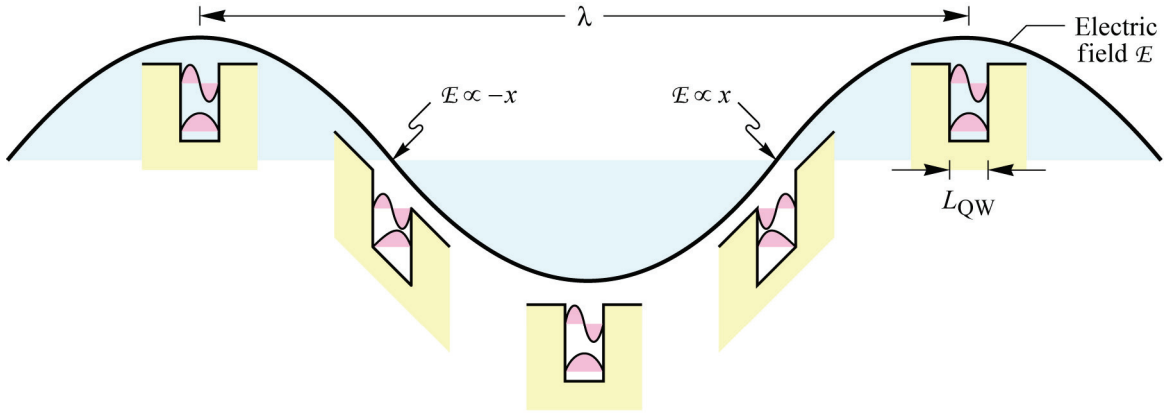


Fig. 11.6. Perturbation of a quantum-well potential by the electric field of an electromagnetic wave with wavelength λ. The linear part of the electric field, where the potential perturbation is $U(x) = -e\mathcal{E}_0 x$, has the greatest perturbative effect on the wave function. In contrast, a constant-potential perturbation just "lifts up" the quantum well with the wave functions not changing in shape or symmetry. Furthermore, for $\lambda >> L_{QW}$, the perturbing electric field always can be approximated by a field that depends linearly on x.

Consider a semiconductor illuminated by an electromagnetic field polarized along the x direction. The electromagnetic field, with angular frequency ω at a particular position, is given by

$$\vec{\mathcal{E}}(t) = \mathcal{E}_0 \, \vec{\mathbf{u}}_x \cos \omega t \tag{11.41}$$

where $\vec{\mathbf{u}}_x$ is the unit vector along the x direction and $\mathcal{E}_0$ is the amplitude of the electric field. As shown in ***Fig***. 11.6, the perturbing potential energy caused by an oscillating electromagnetic

field (such as a light wave) is given by

$$U(x) \quad = \quad -e\,\mathcal{E}_0\,x \tag{11.42}$$

Thus the perturbation hamiltonian of an electron subjected to an electric field is given by

$$H'(x) \quad = \quad -e\,\mathcal{E}_0\,x\,\frac{1}{2}\left(e^{i\omega t} + e^{-i\omega t}\right) \tag{11.43}$$

The symmetry of the perturbation hamiltonian, and the symmetries of the initial and final wave functions give rise to ***selection rules*** that *allow* or *disallow* certain transitions. ***Fig***. 11.7 summarizes the selection rules for different optical transitions in semiconductors.

(1) *Interband transitions and intraband transitions* in bulk semiconductor structures occur *between* bands and *within* a band, respectively. Show that the transition matrix element for interband (band-to-band) stimulated absorption is non-zero. Show that the transition matrix element for intraband stimulated absorption is zero.

(2) *Interband transitions* in quantum well structures occur between quantized states in the conduction band well and quantized states in the valence band well. Let us assign these transitions an energy E_{mn}, where m is the mth quantized state in the conduction band well and n is the nth quantized state in the valence band well. For example, the E_{00} transition occurs between the two ground states of the wells. Show that the transitions E_{00}, E_{02}, E_{20} … are allowed transitions and that E_{01}, E_{10}, E_{12} … are disallowed transitions. Show that allowed interband transitions are characterized by $\Delta = m - n = 0, 2, 4 \ldots$.

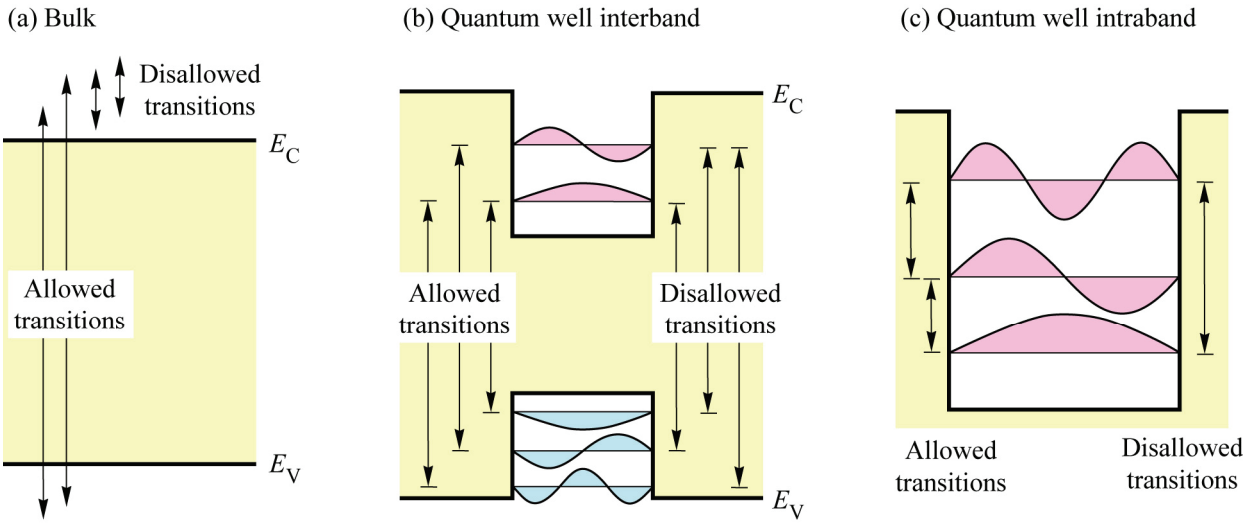


Fig. 11.7. Allowed and disallowed optical interband and intraband transitions in bulk and quantum well semiconductors.

(3) *Intraband transitions* in quantum well structures occur between quantized states in the same well. Such intraband transitions typically occur in the far infrared. Thermally sensitive cameras are based on this principle. Let us assign these transitions an energy E_{mn}, where m is the mth quantized state in the well and n is the nth quantized state in the well. For example, the E_{10} transition occurs between the first excited state and the ground state of the well. Show that the

transitions E_{01}, E_{03}, E_{12} ... are allowed transitions and that E_{02}, E_{04}, E_{24} ... are disallowed transitions. Show that allowed intraband transitions are characterized by $\Delta = m - n = 1, 3, 5 \ldots$.

Discuss how the selection rules for quantum well interband and intraband transitions change in the presence of a static electric field.

Solution: In the absence of an electric field, wave functions in a quantum well are perfectly symmetric or anti-symmetric. As a result, some transitions are allowed and some are forbidden. In the presence of an electric field, wave functions in a quantum well (or other structure) no longer are perfectly symmetric or anti-symmetric. As a result, forbidden transitions become allowed.

Exercise 2: ***Einstein's $\mathcal{A}$ and $\mathcal{B}$ coefficients***. The first theory of spontaneous and stimulated transitions was developed by Einstein. He used the coefficients $\mathcal{A}$ and $\mathcal{B}$ to characterize stimulated and spontaneous transitions in an atom with two quantized levels. Denoting the two levels as "1" and "2", Einstein postulated the probability for the downward transition ($2 \rightarrow 1$) and upward transition ($1 \rightarrow 2$) as

$$W_{2\rightarrow1} = \mathcal{B}_{2\rightarrow1}\,\rho(\nu) + \mathcal{A} \tag{11.44}$$

$$W_{1\rightarrow2} = \mathcal{B}_{1\rightarrow2}\,\rho(\nu) \tag{11.45}$$

The downward transition probability (per atom) is the sum of an induced term $\mathcal{B}_{2\rightarrow1}\,\rho(\nu)$ proportional to the radiation density $\rho(\nu)$, and a spontaneous term $\mathcal{A}$. The upward probability is just $\mathcal{B}_{1\rightarrow2}\,\rho(\nu)$. Using thermal equilibrium considerations, Einstein showed that $\mathcal{B} = \mathcal{B}_{2\rightarrow1} = \mathcal{B}_{1\rightarrow2}$. Thus stimulated absorption and stimulated emission are truly complementary processes. Einstein also showed that the ratio of the coefficients at a frequency ν in an isotropic medium with refractive index $\bar{n}$ is a constant given by $\mathcal{A} / \mathcal{B} = 8\,\pi\,\bar{n}^{\,3}\,h\,\nu^3 / c^3$.

The equivalence of $\mathcal{B}_{2\rightarrow1}$ and $\mathcal{B}_{1\rightarrow2}$ can be shown by quantum mechanical considerations. Apply Fermi's Golden Rule to a two-level system and show that $\mathcal{B}_{2\rightarrow1} = \mathcal{B}_{1\rightarrow2}$.

Solution: Fermi's Golden Rule is formally identical for excitation and de-excitation transitions. Thus an optical absorption transition ($1 \rightarrow 2$) is formally equivalent to an optical stimulated emission transition ($2 \rightarrow 1$) and the constants that govern this transition must be the same, i.e. $\mathcal{B}_{2\rightarrow1} = \mathcal{B}_{1\rightarrow2}$.

Exercise 3: ***Maximum gain and absorption in semiconductors.*** Show that the magnitude of optical gain in a semiconductor at a given energy cannot exceed the magnitude of the absorption coefficient.

Solution: Maximum absorption is obtained if all conduction band states are empty and the valence band is completely filled. Similarly, maximum gain is obtained if all conduction band states are filled and the valence band is empty. Fermi's Golden Rule and the equivalence of Einstein's $\mathcal{B}$ coefficients for upward and downward transitions show that the probability for the two processes is the same. Thus stimulated emission probability in a

population-inverted semiconductor is equal to the stimulated absorption probability in a non-inverted semiconductor.

The absorption coefficient in GaAs near the band edge is $\alpha \approx 10^4$ cm^{-1}. Typical gains in GaAs are lower than that, typically on the order of 10^2 to 10^3 cm^{-1} due to other loss mechanisms, *e. g.* light scattering at imperfections or waveguide losses.

Exercise 4: ***Spontaneous emission enhancement in a microcavity***. In a low-loss Fabry–Perot resonator, the electric field intensity *inside* the cavity is different from the electric field intensity *outside* the cavity. If the electromagnetic field has a cavity-resonant frequency, the difference is given by the factor F, the cavity finesse. For a low-loss optical cavity consisting of two coplanar reflectors with reflectivities R_1 and R_2, the finesse of is given by

$$F = \frac{\sqrt[4]{R_1 R_2}}{1 - \sqrt{R_1 R_2}} \tag{11.46}$$

Give an intuitive explanation as to why the electric field is higher inside the cavity than outside the cavity as shown in ***Fig***. 11.8. Show that the absorption probability and stimulated emission probability of an atom located inside a cavity is a factor of F higher than if the atom were outside the cavity.

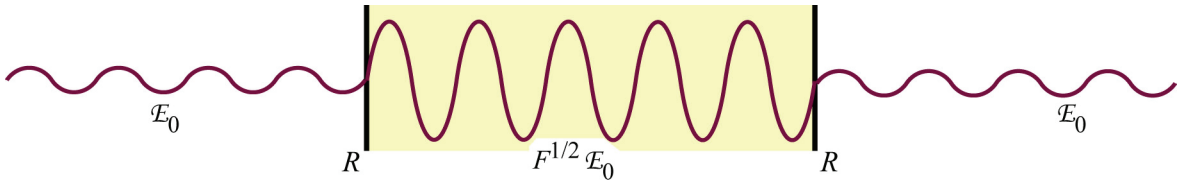


Fig. 11.8. Magnitude of an electric field inside ($F^{1/2}\,\mathcal{E}_0$) and outside ($\mathcal{E}_0$) a symmetric Fabry-Perot resonator with two reflectors with reflectivity R. The frequency of the electric field is assumed to be equal to the resonance frequency of the cavity.

Solution: Due to the high reflectivities of the mirrors, light is trapped inside the cavity and bounces back and forth many times before laving the cavity. As a result, both the light intensity and the electric field strength are much higher inside the cavity than it is outside the cavity.

The probability of a stimulated emission process is proportional to the magnitude of the stimulus, i.e. proportional to the electric field. In Einstein's model, the ratio between the $\mathcal{A}$ and $\mathcal{B}$ coefficients is a constant, *i. e.*, $(\mathcal{A}/\mathcal{B}) = \text{constant}$, where $\mathcal{A}$ and $\mathcal{B}$ describe the spontaneous and stimulated emission probability, respectively. Therefore, the *spontaneous* and *stimulated emission* probability of a semiconductor located in a microcavity is enhanced by a factor given by the cavity finesse F. This principle is used in *resonant-cavity devices*, in which an optical microcavity is integrated with an optoelectronic device, such as a light-emitting diode or a detector.

Enrico Fermi (1901–1954)
Developed “Golden Rule” of time-dependent perturbation theory

12

Density of states

The concentration of neutral impurities, ionized impurities, and free carriers in a doped semiconductor depends on a large number of parameters such as the impurity atom concentration, the free carrier mass, the bandgap energy, and the dielectric constant. The interdependences of the free majority and minority carrier concentration, the impurity concentration, impurity ionization energy as well as some other constants and materials parameters are given by ***semiconductor statistics***. Semiconductor statistics describes the probabilities that a set of electronic states are either vacant or populated.

Electronic states include localized impurity states as well as delocalized conduction and valence band states. In the simplest case, an impurity has a single state with no degeneracy ($g_0 = 1$). However, an impurity may have a degenerate ground state ($g_0 > 1$) as well as excited levels which may need to be considered. The states in the bands and their dependence on energy are described by the ***density of states***. In semiconductor heterostructures, the free motion of carriers is restricted to two, one, or zero spatial dimensions. In order to apply semiconductor statistics to such systems of reduced dimensions, the density of states in quantum wells (two dimensions), quantum wires (one dimension), and quantum dots (zero dimensions), must be known. The density of states in such systems will also be calculated in this chapter.

12.1 Density of states in bulk semiconductors (3D)

Carriers occupy either localized impurity states or delocalized continuum states in the conduction band or valence band. In the simplest case, each impurity has a single, non-degenerate state. Thus, the density of impurity states equals the concentration of impurities. The energy of the impurity states is the same for all impurities (of the same species) as long as the impurities are sufficiently far apart and do not couple. The density of continuum states is more complicated and will be calculated in the following sections. Several cases will be considered including (*i*) a spherical, single-valley band, (*ii*) an anisotropic band, (*iii*) a band with multiple valleys, and (*iv*) the density of states in a semiconductor with reduced degrees of freedom such as quantum wells, quantum wires, and quantum boxes. Finally the *effective* density of states will be calculated.

Single-valley, spherical, and parabolic band

The simplest band structure of a semiconductor consists of a single valley with an isotropic (*i. e*. spherical), parabolic dispersion relation. This situation is closely approximated by, for example, the conduction band of GaAs. The electronic density of states is defined as the number of electron states per unit volume and per unit energy. The finiteness of the density of states is a result of the ***Pauli principle***, which states that only two electrons of opposite spin can occupy one volume element in phase space. The ***phase space*** is defined as a six-dimensional space composed of real space and momentum space. We now define a 'volume' element in phase space to consist of a range of positions and momenta of a particle, such that the position and momentum of the particle are distinguishable from the positions and momenta of other particles.

In order to be distinguishable, the range of positions and momenta must be equal or exceed the range given by the ***uncertainty relation***. The volume element in phase space is then given by

$$\Delta x\, \Delta y\, \Delta z\, \Delta p_x\, \Delta p_y\, \Delta p_z \quad = \quad (2\pi\hbar)^3 \ . \tag{12.1}$$

The 'volume' element in phase space is $(2\pi\hbar)^3$. For systems with only one degree of freedom, Eq. (12.1) reduces to the one-dimensional Heisenberg uncertainty principle $\Delta x\, \Delta p_x = 2\pi\hbar$. The Pauli principle states that two electrons of opposite spin occupy a 'volume' of $(2\pi\hbar)^3$ in phase space. Using the de Broglie relation ($p = \hbar k$) the 'volume' of phase space can be written as

$$\Delta x\, \Delta y\, \Delta z\, \Delta k_x\, \Delta k_y\, \Delta k_z \quad = \quad (2\pi)^3 \ . \tag{12.2}$$

The ***density of states*** per unit energy and per unit volume, which is denoted by $\rho_{DOS}(E)$, allows us to determine the total number of states per unit volume in an energy band with energies E_1 (bottom of band) and E_2 (top of band) according to

$$N \quad = \quad \int_{E_1}^{E_2} \rho_{DOS}(E)\, dE \ . \tag{12.3}$$

Note that N is the total number of states per unit volume, and $\rho_{DOS}(E)$ is the density of states per unit energy per unit volume. To obtain the density of states per unit energy dE, we have to determine how much unit-volumes of k-space is contained in the energy interval E and $E + dE$, since we already know that one unit volume of k-space can contain two electrons of opposite spin.

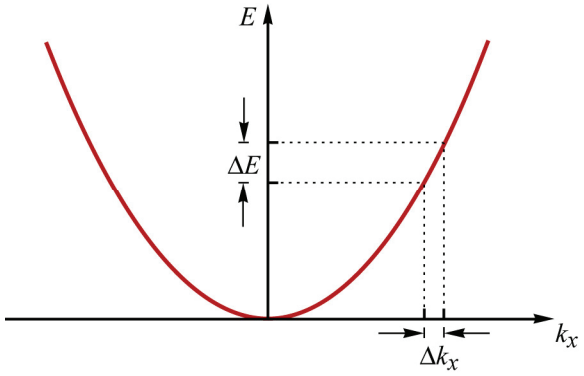


Fig. 12.1. Parabolic dispersion relation with a k-space interval Δk_x and a corresponding energy interval ΔE, where ΔE and Δk_x are related by:

$$\Delta E = (\partial E/\partial k_x)\, \Delta k_x$$

and

$$\Delta k_x = (\partial k_x/\partial E)\, \Delta E.$$

In order to obtain the volume of k-space included between two energies, the *dispersion relation* will be employed. A one-dimensional, parabolic dispersion relation $E = E(k_x)$ is shown in ***Fig***. 12.1. For a given dE one can easily determine the corresponding length in k-space, as illustrated in ***Fig***. 12.1. The k-space length associated with an energy interval dE is simply given by the slope of the dispersion relation. While the one-dimensional dispersion relations can be illustrated easily, the three-dimensional dispersion relation cannot be illustrated in three-dimensional space. To circumvent this difficulty, *surfaces of constant energy in k-space* are frequently used to illustrate a three-dimensional dispersion relation. As an example, the constant energy surface in k-space is illustrated in ***Fig***. 12.2 for a spherical, single-valley band. A large separation of the constant energy surfaces, *i. e.* a large Δk for a given ΔE, indicates a weakly

curved dispersion and a large effective mass.

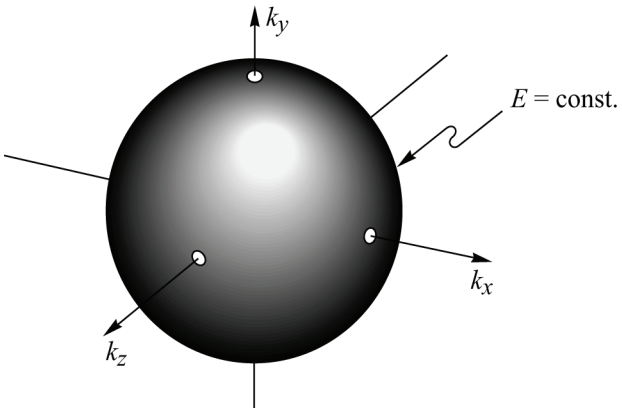


Fig. 12.2. Constant energy surface for a single-valley, isotropic band.

In order to obtain the volume of k-space enclosed between two constant energy surfaces, which correspond to energies E and $E + \mathrm{d}E$, we (first) determine $\mathrm{d}k$ associated with $\mathrm{d}E$ and (second) integrate over the entire constant energy surface. The 'volume' of k-space enclosed between the two constant energy surfaces shown in ***Fig***. 12.3 is thus given by

$$V_{k-\mathrm{space}}(E) \quad = \quad \mathrm{d}E \int_{\mathrm{Surface}} \frac{\partial\, k}{\partial\, E(k)}\, \mathrm{d}s \tag{12.4}$$

where $\mathrm{d}s$ is an area element of the constant energy surface. In a three-dimensional k-space we use $\mathrm{grad}_k = (\partial / \partial k_x, \partial / \mathrm{d}k_y, \partial / \partial k_z)$ and obtain

$$V_{k-\mathrm{space}}(E) \quad = \quad \mathrm{d}E \int_{\mathrm{Surface}} \frac{\mathrm{d}s}{\nabla_k\, E(k)}\ . \tag{12.5}$$

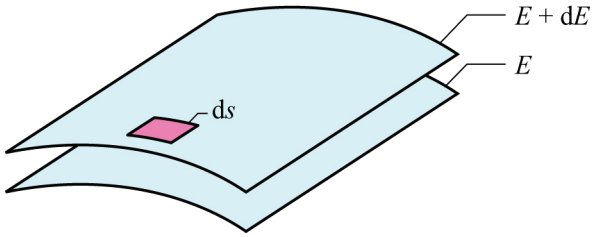


Fig. 12.3. Constant-energy surfaces with energy E and $E + \mathrm{d}E$ used to calculate volume in k-space enclosed betweeen the two surfaces.

Since an electron requires a volume of $4\pi^3$ in phase space, the number of states per unit volume is given by

$$N(E) \quad = \quad \frac{1}{4\pi^3}\ \mathrm{d}E \int_{\mathrm{Surface}} \frac{\mathrm{d}s}{\nabla_k\, E(k)}\ . \tag{12.6}$$

Finally, we obtain the density of states per unit energy and unit volume according to

$$\rho_{DOS}(E) \;=\; \frac{1}{4\pi^3}\int_{Surface}\frac{ds}{\nabla_k\, E(k)}\;. \tag{12.7}$$

In this equation, the surface element ds is always perpendicular to the vector grad$_k$ E (k). Note that the surface element ds is in k-space and that ds has the dimension m^{-2}.

Next we apply the expression for the density of states to *isotropic parabolic* dispersion relations of a three-dimensional semiconductor. In this case the surface of constant energy is a sphere of area $4\pi k^2$ and the parabolic dispersion is $E = \hbar^2 k^2/(2m^*) + E_{pot}$ where k is the wave vector. Insertion of the dispersion in Eq. (12.7) yields the density of states in a semiconductor with a single-valley, isotropic, and parabolic band

$$\boxed{\rho_{DOS}^{3D}(E) \;=\; \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E-E_{pot}}} \tag{12.8}$$

where E_{pot} is a potential energy such as the conduction band edge or the valence band edge energy, E_C or E_V, respectively.

Exercise 1: ***Three-dimensional density of states***. Derive the equation for the 3D density of states, Eq. (12.8).

Solution: From Eq. (12.7), the density of states per unit energy and unit volume is,

$$\rho_{DOS}^{3D}(E) \;=\; \frac{1}{4\pi^3}\int_{Surface}\frac{ds}{\nabla_k\, E(k)}$$

Using $k^2 = k_x^2 + k_y^2 + k_z^2$ and $E = \hbar^2(k_x^2 + k_y^2 + k_z^2)/(2m^*) + E_{pot}$, one obtains

$$\nabla_k\, E(k) \;=\; \hbar^2/(2m^*)\;(2k_x, 2k_y, 2k_z) = (\hbar^2/m^*)\;(k_x, k_y, k_z)$$

In momentum-space, the direction of (k_x, k_y, k_z) is same as the surface-normal vector. Therefore, the integration over the sphere's surface area yields,

$$\int_{Surface}\frac{ds}{\nabla_k\, E(k)} \;=\; \left|\frac{1}{\nabla_k\, E(k)}\right| 4\pi k^2 \;=\; 4\pi k^2\,\frac{1}{(\hbar^2/m^*)k} \;=\; \frac{4\pi m^* k}{\hbar^2}$$

where $4\pi k^2$ is the surface area of the constant-energy surface, i.e. a sphere. From Eq. (12.7),

$$\rho_{DOS}^{3D}(E) \;=\; \frac{1}{4\pi^3}\int_{Surface}\frac{ds}{\nabla_k E(k)} \;=\; \frac{1}{4\pi^3}\,\frac{4\pi m^* k}{\hbar^2} \;=\; \frac{m^* k}{\pi^2\hbar^2}$$

From E-versus-k relationship, i.e. $E - E_{pot} = \hbar^2 k^2/(2m^*)$, we have

$$k = \sqrt{2m^*(E-E_{pot})}/\hbar$$

Substituting k into $\rho_{DOS}{}^{3D}(E)$ yields

$$\rho_{\mathrm{DOS}}^{3\mathrm{D}}(E) = \frac{km^*}{\pi^2\hbar^2} = \frac{m^*}{\pi^2\hbar^2}\sqrt{\frac{2m^*(E-E_{\mathrm{pot}})}{\hbar^2}} = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E-E_{\mathrm{pot}}}\ ,$$

what was to be shown.

Single-valley, anisotropic, parabolic band

In an anisotropic single-valley band, the dispersion relation depends on the spatial direction. Such an anisotropic dispersion is found in III–V semiconductors in which the L- or X- point of the Brillouin zone is the lowest minimum, for example in GaP or AlAs. The surface of constant energy is then no longer a sphere, but an ellipsoid, as shown in ***Fig***. 12.4. The three main axes of the ellipsoid may have different lengths, and thus the three dispersion relations are curved differently. If the main axes of the ellipsoid align with a cartesian coordinate system, the dispersion relation is

$$E = \frac{\hbar^2 k_x^2}{2 m_x^*} + \frac{\hbar^2 k_y^2}{2 m_y^*} + \frac{\hbar^2 k_z^2}{2 m_z^*}\ . \tag{12.9}$$

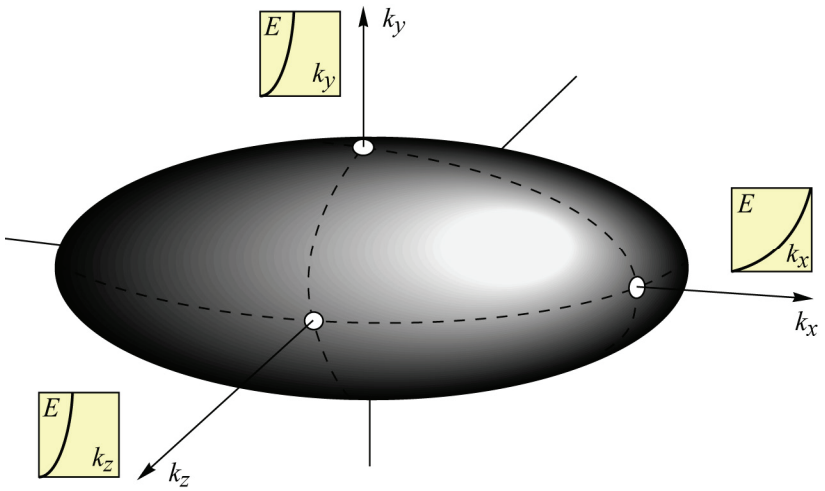


Fig. 12.4. Ellipsoidal constant energy surface with a weakly curved dispersion relation along the k_x axis and strongly curved dispersion along the k_y and k_z axis.

The vector $\mathrm{grad}_k E$ is given by $\mathrm{grad}_k\ E = (\hbar^2 k_x / m_x^*,\ \hbar^2 k_y / m_y^*,\ \hbar^2 k_z / m_z^*)$. Since the vector $\mathrm{grad}_k E$ is perpendicular on the surface element, the *absolute* values of ds and $\mathrm{grad}_k E$ can be taken for the integration. Integration of Eq. (12.7) with the dispersion relation of Eq. (12.9) yields the density of states in an anisotropic semiconductor with parabolic dispersion relations, *i. e.*

$$\rho_{\mathrm{DOS}}(E) = \frac{\sqrt{2}}{\pi^2\hbar^3}\sqrt{m_x^* m_y^* m_z^*}\sqrt{E-E_{\mathrm{pot}}}\ . \tag{12.10}$$

If the main axes of the constant-energy ellipsoid do not align with the k_x, k_y, and k_z axes of the coordinate system then m_x^*, m_y^*, and m_z^* can be formally replaced by m_1^*, m_2^*, and m_3^*.

Frequently, the constant energy surfaces are rotational ellipsoids, that is, two of the main

axes of the ellipsoid are identical. The axes are then denoted as the transversal and the longitudinal axes for the short and long axes, respectively. Such a rotational ellipsoid is schematically shown in ***Fig***. 12.4. A relatively light mass is associated with the (short) transversal axis, while a relatively heavy mass is associated with the (long) longitudinal axis. If the masses are denoted as m_t* and m_l* for the transversal and the longitudinal mass, respectively, Eq. (12.10) can be modified according to

$$\rho_{DOS}(E) = \frac{\sqrt{2}}{\pi^2 \hbar^3} \sqrt{m_l^* \, m_t^{*2}} \sqrt{E - E_{pot}} \quad . \tag{12.11}$$

The anisotropic masses m_x*, m_y*, m_z*, m_l*, and m_t* are frequently used to define a ***density-of-states effective mass***. This mass is given by

$$\boxed{m_{DOS}^* = (m_x^* \, m_y^* \, m_z^*)^{1/3}} \tag{12.12a}$$

$$\boxed{m_{DOS}^* = (m_t^2 \, m_l)^{1/3}} \tag{12.12b}$$

The density of states is then given by

$$\rho_{DOS}(E) = \frac{\sqrt{2}}{\pi^2 \hbar^3} (m_{DOS}^*)^{3/2} \sqrt{E - E_{pot}} \quad . \tag{12.13}$$

Note that for isotropic semiconductors the effective mass coincides with the density-of-states effective mass.

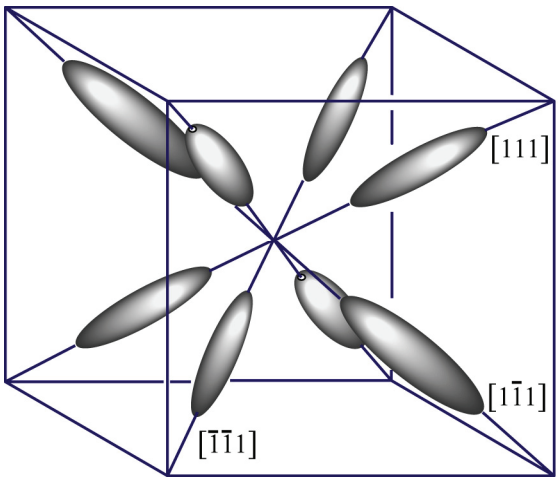


Fig. 12.5. Constant energy surface for the *L*-point of the Brillouin zone. The band structure consists of eight equivalent rotational ellipsoids.

Multiple valleys

At several points of the Brillouin zone, several equivalent minima occur. For example, eight equivalent minima occur at the L-point as schematically shown in ***Fig***. 12.5. Each of the valleys can accommodate carriers, since the minima occur at different k_x, k_y, and k_z values, *i. e.* the Pauli principle is not violated. The density of states is thus obtained by multiplication with the number of equivalent minima, that is

$$\rho_{DOS}(E) = \frac{M_c \sqrt{2}}{\pi^2 \hbar^3} \sqrt{m_1^* m_2^* m_3^*} \sqrt{E - E_{pot}} \tag{12.14}$$

where M_c is the number of equivalent minima and m_1*, m_2*, and m_3* are the effective masses for motion along the three main axes of the ellipsoid.

12.2 Density of states in semiconductors with reduced dimensionality (2D, 1D, 0D)

Semiconductor heterostructure allows one to change the band energies in a controlled way and confine charge carriers to two (2D), one (1D), or zero (0D) spatial dimensions. Due to the confinement of carriers, the dispersion relation along the confinement direction is changed. The change in dispersion relation results in a change in the density of states.

Confinement of a carrier in one spatial dimension, *e. g.* the *z*-direction results in the formation of quantum states for motion along this direction. Consider the ground state in a quantum well of width L_z with infinitely high walls. The ground-state energy is obtained from the solution of Schrödinger's equation and is given by

$$E_0 = \frac{\hbar^2}{2 m^*} \left(\frac{\pi}{L_z} \right)^2 . \tag{12.15}$$

The particle in the quantum well can assume a range of momenta in the *z*-direction; the range is given by the uncertainty principle, *i. e.*

$$\Delta k_z = \frac{\Delta p_z}{\hbar} = \frac{2\pi}{L_z} . \tag{12.16}$$

The dispersion relation for motion along the confinement (*z*-) direction is thus given by

$$E = E_0 \qquad \text{for entire range of } k_z . \tag{12.17}$$

The dispersion is flat, *i. e.* constant for all values of k_z. The *z*-component of the vector $\text{grad}_k E$ (see Eq. 12.7) is therefore zero and need not be considered.

We next consider the *x*- and *y*- direction and recall that the Schrödinger equation is separable for the three spatial dimensions. Thus, the kinetic energy in the xy-plane is given by

$$E = \frac{\hbar^2}{2 m^*} \left(k_x^2 + k_y^2 \right) \tag{12.18}$$

for a parabolic dispersion.

The surface of constant energy for the dispersion relation given by Eq. (12.18) is shown in ***Fig***. 12.6, and is a circle around $k_x = k_y = 0$. The density of states of such a 2D electron system is obtained by similar considerations as for the 3D case. The reduced phase space now consists only of the xy-plane and the k_x and k_y coordinates. Correspondingly, the two-dimensional density of states is the number of states per *unit-area* and unit-energy. The volume of *k*-space between the circles of constant energy is given by Eq. (12.5). The equation is evaluated most conveniently in polar coordinates in which $k_r = (k_x^2 + k_y^2)^{1/2}$ is the radial component of the *k*-vector. The surface integral reduces to a line integral and the total length of the circular line is

$2\pi k_r$. The volume of k-space then obtained is

$$V_{k-\text{space}}^{2D}(E) = dE \int_{\text{Surface}} \frac{ds}{\nabla_k E(k)} = \frac{2\pi m^*}{\hbar^2} . \qquad (12.19)$$

Since two electrons of opposite spin require a volume element of $(2\pi)^2$ in phase space, the density of states of a 2D electron system is given by

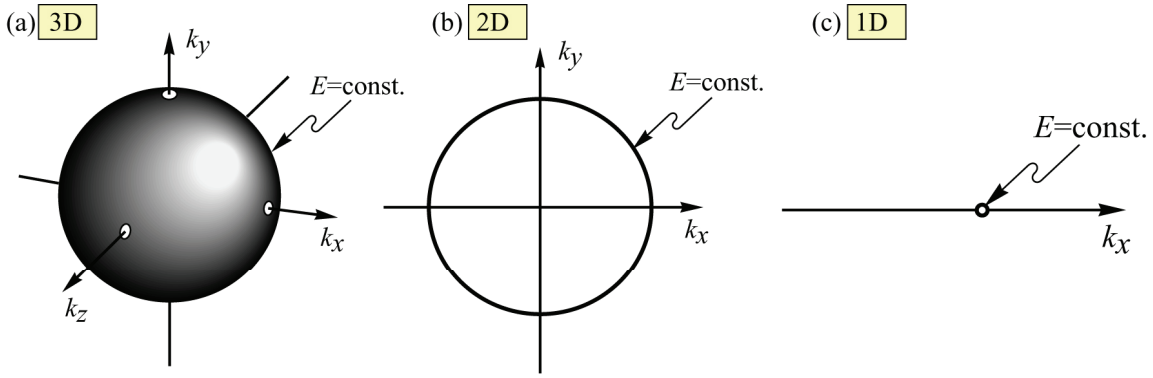


Fig. 12.6. Constant energy surfaces of a (a) 3-dimensional, (b) 2-dimensional, and (c) 1-dimensional system. The surfaces are a sphere, a circle, and a point for 3D, 2D, and 1D systems, respectively.

$$\boxed{\rho_{DOS}^{2D}(E) = \frac{m^*}{\pi\hbar^2}} \qquad (E \geq E_0) \qquad (12.20)$$

where E_0 is the ground state of the quantum well system. For energies $E \geq E_0$, the 2D density of states is a constant and does not depend on energy. If the 2D semiconductor has more than one quantum state, each quantum state has a state density of Eq. (12.20). The total density of states can be written as

$$\rho_{DOS}^{2D}(E) = \frac{m^*}{\pi\hbar^2} \sum_n \sigma(E - E_n) \qquad (12.21)$$

where E_n are the energies of quantized states and $\sigma(E - E_n)$ is the step function.

Exercise 2: ***Two-dimensional density of states***. Derive the equation for the 2D density of states, Eq. (12.20).

Solution: According to Eq. (12.19), the volume in k-space is given by

$$V_{k-\text{space}}^{2D}(E) = dE \int_{\text{Surface}} \frac{ds}{\nabla_k E(k)}$$

Each two electrons with opposite spin require a k-space volume of $(2\pi)^2$. Therefore, the density of states per unit energy per unit area is given by

$$\rho_{\mathrm{DOS}}^{\mathrm{2D}}(E) \;=\; \frac{2}{(2\pi)^2}\int_{\mathrm{Surface}} \frac{\mathrm{d}s}{\nabla_k\, E(k)}\;.$$

The electron-energy-versus-momentum relationship is given by $E = \hbar^2 k^2 / (2m^*) + E_{\mathrm{pot}}$. In two-dimensional space, the momentum has two components only, so that $k = k_x^2 + k_y^2$. Therefore, $E = \hbar^2 (k_x^2 + k_y^2) / (2m^*) + E_{\mathrm{pot}}$ and

$$\nabla_k\, E(k) \;=\; \hbar^2/(2m^*)\;(2k_{\mathrm{x}}, 2k_{\mathrm{y}}) \;=\; (\hbar^2/m^*)\;(k_{\mathrm{x}}, k_{\mathrm{y}})$$

In 2D momentum-space, the direction of the vector (k_x, k_y) is the same as the surface-normal vector. Therefore,

$$\int_{\mathrm{Surface}} \frac{\mathrm{d}s}{\nabla_k\, E(k)} \;=\; \left|\frac{1}{\nabla_k\, E(k)}\right| 2\pi k \;=\; 2\pi k\,\frac{1}{(\hbar^2/m^*)\,k} \;=\; \frac{2\pi m^*}{\hbar^2}$$

where $2\pi k$ is the circumference of the constant-energy surface, i.e. a circle. Hence,

$$\rho_{\mathrm{DOS}}^{\mathrm{2D}}(E) \;=\; \frac{2}{(2\pi)^2}\int_{\mathrm{Surface}} \frac{\mathrm{d}s}{\nabla_k E(k)} \;=\; \frac{2}{(2\pi)^2}\,\frac{2\pi m^*}{\hbar^2} \;=\; \frac{m^*}{\pi\hbar^2}\;,$$

what was to be shown.

We next consider a one-dimensional (1D) system, the quantum wire, in which only one direction of motion is allowed, *e. g.* along the x-direction. The dispersion relation is then given by $E = \hbar^2 k_x^2/(2m^*)$. The 'volume' (*i. e.* length-unit) in k-space is obtained in analogy to the three-dimensional and two-dimensional case according to Eq. (12.5). The 'surface' integral reduces to a single point in k-space, *i. e.* the point $k = k_x$. Thus, the volume of k-space is given by

$$V_{k-\mathrm{space}}^{\mathrm{1D}}(E) \;=\; \int_{\mathrm{Surface}} \frac{\delta(k_x - k_{x0})\;\mathrm{d}s}{\nabla_k\, E(k_x)} \;=\; \sqrt{\frac{m^*}{2\,\hbar^2 (E - E_0)}} \qquad (E \geq E_0)\,. \qquad (12.22)$$

The volume in phase space of two electrons with opposite spin is given by 2π and thus the 1D density of states is given by

$$\boxed{\rho_{\mathrm{DOS}}^{\mathrm{1D}}(E) \;=\; \frac{1}{\pi\hbar}\sqrt{\frac{m^*}{2\,(E - E_0)}}} \qquad (E \geq E_0)\,. \qquad (12.23)$$

Note that the density of states in a 3-, 2- and 1-dimensional system has a functional dependence on energy according to $E^{1/2}$, E^0, and $E^{-1/2}$, respectively. For more than one quantized state, the 1D density of states is given by

$$\rho_{\mathrm{DOS}}^{\mathrm{1D}}(E) \;=\; \frac{1}{\pi\hbar}\;\sum_n \sqrt{\frac{m^*}{2\,(E - E_n)}}\;\sigma(E - E_n) \qquad (12.24)$$

where E_n are the energies of the quantized states of the wire.

Finally, we consider the density of states in a zero-dimensional (0D) system, the quantum

box. No free motion is possible in such a quantum box, since the electron is confined in all three spatial dimensions. Consequently, there is no *k*-space available which could be filled up with electrons. Each quantum state of a 0D system can therefore be occupied by only two electrons. The density of states is therefore described by a δ-function.

$$\rho_{DOS}^{0D}(E) = 2\,\delta(E - E_0) \tag{12.25}$$

For more than one quantum state, the density of states is given by

$$\rho_{DOS}^{0D}(E) = \sum_n 2\,\delta(E - E_n)\ . \tag{12.26}$$

The densities of states for one quantized level for a 3D, 2D, 1D, and 0D electron system are schematically illustrated in ***Fig***. 12.7.

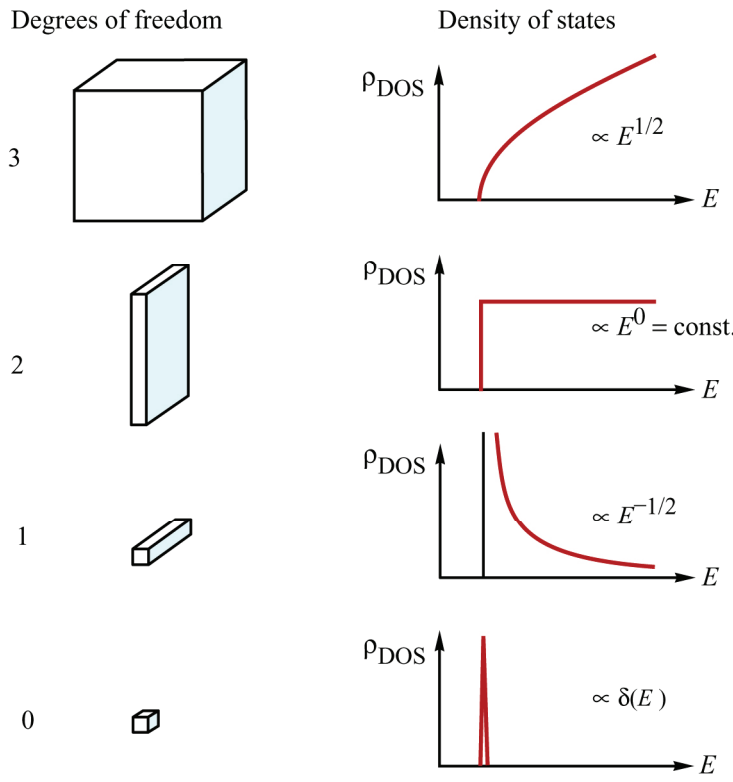


Fig. 12.7. Electronic density of states of semiconductors with 3, 2, 1, and 0 degrees of freedom for electron propagation. Systems with 2, 1, and 0 degrees of freedom are referred to as quantum wells, quantum wires, and quantum boxes, respectively.

12.3 Effective density of states in 3D, 2D, 1D, and 0D semiconductors

The ***effective density of states*** is introduced in order to simplify the calculation of the population of the conduction and valence band. The basic simplification made is that all band states are assumed to be located directly at the band edge. This situation is illustrated in ***Fig***. 12.8 for the conduction band. The 3D density of states has square-root dependence on energy. The effective density of states is δ-function-like and occurs at the bottom of the conduction band.

An electronic state can be either occupied by an electron or unoccupied. Quantum mechanics allows us to attribute to the state a probability of occupation. The total electron concentration in a band is then obtained by integration over the product of state density and the probability that the state is occupied, that is

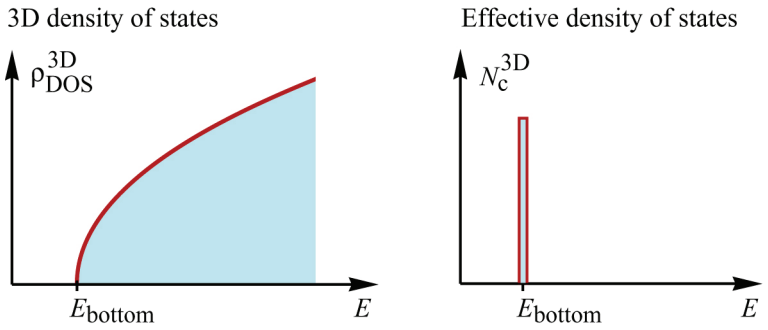


Fig. 12.8. Energy dependent density of states, ρ_{DOS}^{3D}, and *effective* density of states, N_c^{3D}, at the bottom of the conduction band.

$$n = \int_{E_{bottom}}^{E_{top}} \rho_{DOS}(E)\, f(E)\, dE \tag{12.27}$$

where $f(E)$ is the (dimensionless) probability that a state of energy E is populated (see Sect. on *semiconductor statistics*). The limits of the integration are the bottom and the top energy of the band, since the electron concentration in the entire band is of interest.

As will be shown in a subsequent section, the probability of occupation, $f(E)$, is given by the Maxwell–Boltzmann distribution (see Sect. on *Maxwell–Boltzmann distribution*). The Maxwell–Boltzmann distribution, also frequently referred to as the Boltzmann distribution, is given by

$$f_B(E) = \exp\left(-\frac{E - E_F}{k\,T}\right) \tag{12.28}$$

where E_F is the Fermi energy (for a definition of the Fermi energy the reader is again referred to the next section). Using Eq. (12.27), the electron concentration can be determined by evaluating the integral.

The *effective* density of states at the bottom of the conduction band is now defined as the density of states which yields, with the Boltzmann distribution, the *same* electron concentration as the true density of states, that is

$$n = \int_{E_{bottom}}^{E_{top}} \rho_{DOS}(E)\; f_B(E)\; dE = N_c\; f_B(E = E_C) \tag{12.29}$$

where N_c is the effective density of states at the bottom of the conduction band and E_C is the energy of the bottom of this band. Strictly speaking, the effective density of states has no physical meaning but is simply a mathematical tool to facilitate calculations. For completeness, Eqs. (12.27) and (12.29) are now given explicitly using the Boltzmann distribution and the density of states of an isotropic three-dimensional semiconductor:

$$n = \int_{E_C}^{\infty} \frac{1}{2\,\pi^2} \left(\frac{2\,m^*}{\hbar^2}\right)^{3/2} \sqrt{E - E_C}\; e^{-(E-E_F)/kT}\; dE\;, \tag{12.30}$$

$$n = N_c\, e^{-(E_C - E_F)/kT}\;. \tag{12.31}$$

The upper limit of the integration can be taken to be infinity without loss of accuracy due to the strongly converging Boltzmann factor. Evaluation of the integral in Eq. (12.30) and comparison with Eq. (12.31) yields the effective density of states

$$N_c = \frac{1}{\sqrt{2}} \left(\frac{m^* kT}{\pi \hbar^2} \right)^{3/2} \tag{12.32}$$

Note that the effective density of states given by Eq. (12.32) applies to one minimum in the conduction band.

Exercise 3: ***Effective density of states***. Derive the equation for the 3D effective density of states by evaluating Eq. (12.30) and comparing it with Eq. (12.31).

Solution: Equation (12.30) in the text book is

$$n = \int_{E=E_C}^{\infty} \frac{1}{2\pi^2} \left(\frac{2m^*}{\hbar^2} \right)^{3/2} \sqrt{E - E_C} \; e^{-(E-E_F)/kT} \, dE$$

Using the substitution $E^* = E - E_C$ and $dE^* = dE$, the above equation can be written as

$$n = \frac{1}{2\pi^2} \left(\frac{2m^*}{\hbar^2} \right)^{3/2} \int_{E^*=0}^{\infty} \sqrt{E^*} \; e^{-(E^*-(E_F-E_C))/kT} \, dE^* .$$

Using the known mathematical formula

$$\int_0^{\infty} \sqrt{x} \; e^{-\alpha(x-x_0)} \, dx = \frac{1}{2} \frac{e^{\alpha x_0} \sqrt{\pi}}{\alpha^{3/2}}$$

we obtain

$$\int_{E^*=0}^{\infty} \sqrt{E^*} \, e^{-(E^*-(E_F-E_C))/kT} \, dE^* = \frac{1}{2} \frac{e^{(E_F-E_C)/kT} \sqrt{\pi}}{(1/kT)^{3/2}} .$$

Therefore we have,

$$n = \frac{1}{2\pi^2} \left(\frac{2m^*}{\hbar^2} \right)^{3/2} \times \frac{1}{2} \frac{e^{(E_F-E_C)/kT} \sqrt{\pi}}{(1/kT)^{3/2}} = \underbrace{\frac{1}{\sqrt{2}} \left(\frac{m^* kT}{\pi \hbar^2} \right)^{3/2}}_{N_c} e^{-(E_C-E_F)/kT}$$

Comparing this result with Eq. (12.31), yields $N_c = 2^{-1/2} (m^* \, kT / (\pi \hbar^2))^{3/2}$, what was to be shown.

If there are a number of M_c equivalent minima in the conduction band, the corresponding density of states must be multiplied by M_c. Furthermore, if the band structure is anisotropic, the effective mass m^* must be replaced by the density-of-states effective mass m_{DOS}^*. For a degenerate valence band with heavy and light holes, the effective density of states is the sum of both effective state densities, that is

$$N_\mathrm{v} = \frac{1}{\sqrt{2}}\left(\frac{m^*_\mathrm{hh}\,kT}{\pi\hbar^2}\right)^{3/2} + \frac{1}{\sqrt{2}}\left(\frac{m^*_\mathrm{lh}\,kT}{\pi\hbar^2}\right)^{3/2} . \tag{12.33}$$

The effective density of states in a two-dimensional system (*i. e.* a system with two degrees of freedom) is obtained by the identical procedure as the three-dimensional effective density of states. The equations that are analogue to Eqs. (12.30) and (12.31) then read

$$n^\mathrm{2D} = \int_{E_\mathrm{C}}^{\infty} \frac{m^*}{\pi\hbar^2}\,\mathrm{e}^{-(E-E_\mathrm{F})/kT}\,\mathrm{d}E , \tag{12.34}$$

$$n^\mathrm{2D} = N_\mathrm{c}^\mathrm{2D}\,\mathrm{e}^{-(E_\mathrm{C}-E_\mathrm{F})/kT} \tag{12.35}$$

where N_c^{2D} is the two-dimensional effective density of states. The carrier concentration n^2D represents the number of electrons per unit-area and is also referred to as the 2D density. Evaluation of the integral yields

$$\boxed{N_\mathrm{c}^{2D} = \frac{m^*}{\pi\hbar^2}kT} \tag{12.36}$$

Finally, the effective density of states of a one-dimensional (1D) system is obtained in a similar way. The 1D density, *i. e.* the number of carriers per unit length is given by

$$n^\mathrm{1D} = \int_{E_\mathrm{C}}^{\infty} \frac{1}{\pi\hbar}\sqrt{\frac{m^*}{2\,(E-E_\mathrm{C})}}\,\mathrm{e}^{-(E-E_\mathrm{F})/kT}\,\mathrm{d}E , \tag{12.37}$$

$$n^\mathrm{1D} = N_\mathrm{c}^\mathrm{1D}\,\mathrm{e}^{-(E_\mathrm{C}-E_\mathrm{F})/kT} . \tag{12.38}$$

The one-dimensional effective density of states is obtained as

$$\boxed{N_\mathrm{c}^\mathrm{1D} = \sqrt{\frac{m^*\,kT}{2\,\pi\hbar^2}}} \tag{12.39}$$

The evaluation of a zero-dimensional density of states does not yield a simplification of the carrier-density calculation, since the zero-dimensional density of states is δ-function like. Table 12.1 summarizes the dispersion relation, the density of states, and the effective density of states of semiconductors with various degrees of freedom.

Degrees of freedom	Dispersion (kinetic energy)	Density of states	Effective density of states
3 (bulk)	$E = \frac{\hbar^2}{2m^*}(k_x^2 + k_y^2 + k_z^2)$	$\rho_{DOS}^{3D} = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{\frac{3}{2}}\sqrt{E - E_C}$	$N_c^{3D} = \frac{1}{\sqrt{2}}\left(\frac{m^* kT}{\pi\hbar^2}\right)^{\frac{3}{2}}$
2 (slab)	$E = \frac{\hbar^2}{2m^*}(k_x^2 + k_y^2)$	$\rho_{DOS}^{2D} = \frac{m^*}{\pi\hbar^2}\sigma(E - E_C)$	$N_c^{2D} = \frac{m^*}{\pi\hbar^2}kT$
1 (wire)	$E = \frac{\hbar^2}{2m^*}(k_x^2)$	$\rho_{DOS}^{1D} = \frac{m^*}{\pi\hbar}\sqrt{\frac{m^*}{2(E - E_C)}}$	$N_c^{1D} = \sqrt{\frac{m^* kT}{2\pi\hbar^2}}$
0 (box)	–	$\rho_{DOS}^{0D} = 2\delta(E - E_C)$	$N_c^{0D} = 2$

Table 12.1 Density of states for semiconductor with 3, 2, 1, and 0 degrees of freedom for propagation of electrons. The dispersion relations are assumed to be parabolic. The formulas can be applied to anisotropic semiconductors if the effective mass m^* is replaced by the density-of-states effective mass m_{DOS}^*. If the semiconductor has a number of M_c equivalent minima, the corresponding density of states must be multiplied by M_c. The bottom of the band is denoted as E_C and $\sigma(E)$ is the step-function.

Exercise 4: ***Density of states and effective density of states***. What is the (i) density of states and the (ii) effective density of states at the conduction band edge, E_C, of a semiconductor?

Solution: (i) The density of states at the conduction band of a semiconductor at $E = E_C$ is $\rho_{DOS}^{3D} = 0$. (ii) The effective density of states at the conduction band of a semiconductor at $E = E_C$ is N_c where N_c is given in the above table.

What are the units of the density of states in a 3D, 2D, 1D, and 0D semiconductor?

Solution:	***Dimensionality***	***Units***	***Common units***
	3D	$m^{-3}\,J^{-1}$	$cm^{-3}\,eV^{-1}$
	2D	$m^{-2}\,J^{-1}$	$cm^{-2}\,eV^{-1}$
	1D	$m^{-1}\,J^{-1}$	$cm^{-1}\,eV^{-1}$
	0D	J^{-1}	eV^{-1}

13

Classical and quantum statistics

Classical Maxwell–Boltzmann statistics and quantum mechanical Fermi–Dirac statistics are introduced to calculate the occupancy of states. Special attention is given to analytic approximations of the Fermi–Dirac integral and to its approximate solutions in the non-degenerate and the highly degenerate regime. In addition, some numerical approximations to the Fermi–Dirac integral are summarized.

Semiconductor statistics includes both classical statistics and quantum statistics. Classical or Maxwell–Boltzmann statistics is derived on the basis of purely classical physics arguments. In contrast, quantum statistics takes into account two results of quantum mechanics, namely (*i*) the Pauli exclusion principle which limits the number of electrons occupying a state of energy E and (*ii*) the finiteness of the number of states in an energy interval E and $E + \mathrm{d}E$. The finiteness of states is a result of the Schrödinger equation. In this section, the basic concepts of classical statistics and quantum statistics are derived. The fundamentals of ideal gases and statistical distributions are summarized as well since they are the basis of semiconductor statistics.

13.1 Probability and distribution functions

Consider a large number N of free classical particles such as atoms, molecules or electrons which are kept at a constant temperature T, and which interact only weakly with one another. The energy of a single particle consists of *kinetic energy* due to translational motion and an internal energy for example due to rotations, vibrations, or orbital motions of the particle. In the following we consider particles with only kinetic energy due to translational motion. The particles of the system can assume an energy E, where E can be either a discrete or a continuous variable. If N_i particles out of N particles have an energy between E_i and $E_i + \mathrm{d}E$, the probability of any particle having any energy within the interval E_i and $E_i + \mathrm{d}E$ is given by

$$f(E_i)\,\mathrm{d}E \quad = \quad \frac{N_i}{N} \tag{13.1}$$

where $f(E)$ is the ***energy distribution function*** of a particle system. In statistics, $f(E)$ is frequently called the *probability density function*. The total number of particles is given by

$$\sum\nolimits_i N_i \quad = \quad N \tag{13.2}$$

where the sum is over all possible energy intervals. Thus, the integral over the energy distribution function is

$$\int_0^\infty f(E)\,\mathrm{d}E \quad = \quad \sum_i \frac{N_i}{N} \quad = \quad 1\ . \tag{13.3}$$

In other words, the probability of any particle having an energy between zero and infinity is unity. Distribution functions which obey

$$\int_0^\infty f(E)\,\mathrm{d}E = 1 \tag{13.4}$$

are called ***normalized*** distribution functions.

The ***average energy*** or ***mean energy*** $\overline{E}$ of a single particle is obtained by calculating the total energy and dividing by the number of particles, that is

$$\overline{E} = \frac{1}{N}\sum_i N_i\,E = \int_0^\infty E\,f(E)\,\mathrm{d}E\ . \tag{13.5}$$

In addition to energy distribution functions, velocity distribution functions are valuable. Since only the kinetic translational motion (no rotational motion) is considered, the velocity and energy are related by

$$E = \tfrac{1}{2}\,m\,v^2\ . \tag{13.6}$$

The average velocity and the average energy are related by

$$\overline{E} = \tfrac{1}{2}\,m\,\overline{v^2} \tag{13.7}$$

$$v_{\mathrm{rms}} = \sqrt{\overline{v^2}} \tag{13.8}$$

and is the velocity corresponding to the average energy

$$\overline{E} = \tfrac{1}{2}\,m\,v_{\mathrm{rms}}^2\ . \tag{13.9}$$

In analogy to the energy distribution we assume that N_i particles have a velocity within the interval v_i and $v_i + \mathrm{d}v$. Thus,

$$f(v)\,\mathrm{d}v = \frac{N_i}{N} \tag{13.10}$$

where $f(v)$ is the normalized velocity distribution. Knowing $f(v)$, the following relations allow one to calculate the mean velocity, the mean-square velocity, and the root-mean-square velocity

$$\bar{v} = \int_0^\infty v\,f(v)\,\mathrm{d}v\ , \tag{13.11}$$

$$\overline{v^2} = \int_0^\infty v^2\,f(v)\,\mathrm{d}v\ , \tag{13.12}$$

$$v_{\mathrm{rms}} = \sqrt{\overline{v^2}} = \left[\int_0^\infty v^2\,f(v)\,\mathrm{d}v\right]^{1/2}\ . \tag{13.13}$$

Up to now we have considered the velocity as a scalar. A more specific description of the velocity distribution is obtained by considering each component of the velocity $v = (v_x, v_y, v_z)$. If N_i particles out of N particles have a velocity in the 'volume' element $v_x + dv_x$, $v_y + dv_y$, and $v_z + dv_z$, the distribution function is given by

$$f(v_x, v_y, v_z)\ dv_x\, dv_y\, dv_z \quad = \quad \frac{N_i}{N}\ . \tag{13.14}$$

Since $\Sigma_i\, N_i = N$, the velocity distribution function is normalized, *i. e.*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(v_x, v_y, v_z)\ dv_x\, dv_y\, dv_z \quad = \quad 1\ . \tag{13.15}$$

The average of a specific propagation direction, for example v_x is evaluated in analogy to Eqs. (13.11 – 13). One obtains

$$\overline{v_x} \quad = \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v_x\, f(v_x, v_y, v_z)\ dv_x\, dv_y\, dv_z\ , \tag{13.16}$$

$$\overline{v_x^2} \quad = \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v_x^2\, f(v_x, v_y, v_z)\ dv_x\, dv_y\, dv_z\ , \tag{13.17}$$

$$v_{x,\mathrm{rms}} \quad = \quad \sqrt{\overline{v_x^2}} \quad = \quad \left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v_x^2\, f(v_x, v_y, v_z)\ dv_x\, dv_y\, dv_z\right]^{1/2}\ . \tag{13.18}$$

In a closed system the mean velocities are zero, that is $\overline{v}_x = \overline{v}_y = \overline{v}_z = 0$. However, the mean-square velocities are, just as the energy, not equal to zero.

13.2 Ideal gases of atoms and electrons

The basis of classical semiconductor statistics is ideal gas theory. It is therefore necessary to make a small excursion into this theory. The individual particles in such ideal gases are assumed to interact weakly, that is collisions between atoms or molecules are a relatively seldom event. It is further assumed that there is no interaction between the particles of the gas (such as electrostatic interaction), unless the particles collide. The collisions are assumed to be (*i*) *elastic* (*i. e.* total energy and momentum of the two particles involved in a collision are preserved) and (*ii*) of very short duration.

Ideal gases follow the universal gas equation (see *e. g.* Kittel and Kroemer, 1980)

$$P\,V \quad = \quad R\,T \tag{13.19}$$

where P is the pressure, V the volume of the gas, T its temperature, and R is the universal gas constant. This constant is independent of the species of the gas particles and has a value of $R = 8.314\ \mathrm{J\ K^{-1}\ mol^{-1}}$.

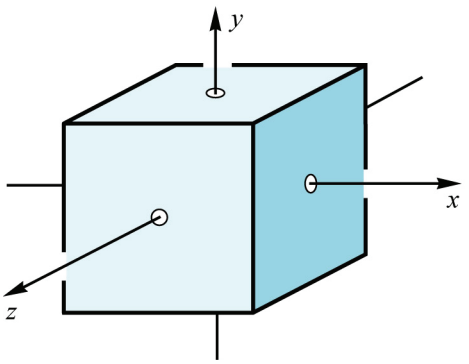


Fig. 13.1. Cubic volume confining one mole (N_{Avo} = 6.023 x 10^{23} atoms/mole) of an ideal gas. The pressure of the ideal gas exerted on one side of the cube (shaded area) is calculated in the text.

Next, the pressure P and the kinetic energy of an individual particle of the gas will be calculated. For the calculation it is assumed that the gas is confined to a cube of volume V, as shown in ***Fig***. 13.1. The quantity of the gas is assumed to be 1 mole, that is the number of atoms or molecules is given by Avogadro's number, $N_{Avo} = 6.023 \times 10^{23}$ particles per mole. Each side of the cube is assumed to have an area $A = V^{2/3}$. If a particle of mass m and momentum $m\,v_x$ (along the x-direction) is elastically reflected from the wall, it provides a momentum $2\,m\,v_x$ to reverse the particle momentum. If the duration of the collision with the wall is dt, then the force acting on the wall during the time dt is given by

$$F = \frac{\mathrm{d}p}{\mathrm{d}t} \tag{13.20}$$

where the momentum change is $\mathrm{d}p = 2\,m\,v_x$. The pressure P on the wall during the collision with one particle is given by

$$\mathrm{d}P = \frac{F}{A} = \frac{1}{A}\frac{\mathrm{d}p}{\mathrm{d}t} \tag{13.21}$$

where A is the area of the cube's walls. Next we calculate the total pressure P experienced by the wall if a number of N_{Avo} particles are within the volume V. For this purpose we first determine the number of collisions with the wall during the time dt. If the particles have a velocity v_x, then the number of particles hitting the wall during dt is $(N_{Avo}/V)\,A\,v_x\,\mathrm{d}t$. The fraction of particles having a velocity v_x is obtained from the velocity distribution function and is given by $f(v_x, v_y, v_z)\,\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z$. Consequently, the total pressure is obtained by integration over all positive velocities in the x-direction

$$P = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{N_{Avo}}{V}\,A\,v_x\,\mathrm{d}t\,f(v_x, v_y, v_z)\;\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z\,\frac{2\,m\,v_x}{A\,\mathrm{d}t}\,. \tag{13.22}$$

Since the velocity distribution is symmetric with respect to positive and negative x-direction, the integration can be expanded from $-\infty$ to $+\infty$.

$$P = \frac{N_{Avo}}{V}\,m\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v_x^2\,f(v_x, v_y, v_z)\;\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z = \frac{N_{Avo}}{V}\,m\,\overline{v_x^2}\,. \tag{13.23}$$

Since the velocity distribution is isotropic, the mean-square velocity is given by

$$\overline{v^2} \;=\; \overline{v_x^2} + \overline{v_y^2} + \overline{v_z^2} \qquad \text{or} \qquad \overline{v_x^2} \;=\; \tfrac{1}{3}\,\overline{v^2}\ . \tag{13.24}$$

The pressure on the wall is then given by

$$P \;=\; \frac{1}{3}\,\overline{v^2}\,\frac{N_{\text{Avo}}}{V}\,m\ . \tag{13.25}$$

Using the universal gas equation, Eq. (13.19), one obtains

$$RT \;=\; \frac{2}{3}\,N_{\text{Avo}}\,\frac{1}{2}\,m\,\overline{v^2}\ . \tag{13.26}$$

The average kinetic energy of one mole of the ideal gas can then be written as

$$\overline{E} \;=\; \overline{E_{kin}} \;=\; \tfrac{3}{2}\,RT\ . \tag{13.27}$$

The average kinetic energy of one single particle is obtained by division by the number of particles, *i. e.*

$$\boxed{\overline{E} \;=\; \overline{E_{\text{kin}}} \;=\; \tfrac{3}{2}\,kT} \tag{13.28}$$

where $k = R / N_{\text{Avo}}$ is the Boltzmann constant. The preceding calculation has been carried out for a three-dimensional space. In a one-dimensional space (one degree of freedom), the average velocity is $\overline{v^2} = \overline{v_x{}^2}$ and the resulting kinetic energy is given by

$$\overline{E_{\text{kin}}} \;=\; \tfrac{1}{2}\,kT \qquad\qquad \text{(per degree of freedom)}\ . \tag{13.29}$$

Thus the kinetic energy of an atom or molecule is given by $(1/2)\,kT$. Equation (13.29) is called the ***equipartition law***, which states that each 'degree of freedom' contributes $(1/2)\,kT$ to the total kinetic energy.

Next we will focus on the energetic distribution of electrons. The properties which have been derived in this section for atomic or molecular gases will be applied to free electrons of effective mass m^* in a crystal. To do so, the interaction between the electrons and the lattice must be negligible and electron – electron collisions must be a relatively seldom event. Under these circumstances we can treat the electron system as a classical ideal gas.

13.3 Maxwell velocity distribution

The Maxwell velocity distribution describes the distribution of velocities of the particles of an ideal gas. It will be shown that the Maxwell velocity distribution is of the form

$$f_{\text{M}}(v) \;=\; A\,\exp\left(-\frac{(1/2)\,m\,v^2}{kT}\right) \tag{13.30}$$

where $(1/2)\,m\,v^2$ is the kinetic energy of the particles. If the energy of the particles is purely

kinetic, the Maxwell distribution can be written as

$$f_M(E) = A \exp\left(-\frac{E}{kT}\right) . \tag{13.31}$$

The proof of the Maxwell distribution of Eq. (13.30) is conveniently done in two steps. In the first step, the exponential factor is demonstrated, *i. e.* $f_M(E) = A \exp(-\alpha E)$. In the second step it is shown that $\alpha = 1/(kT)$.

In the theory of ideal gases it is assumed that collisions between particles are elastic. The total energy of two electrons before and after a collision remains the same, that is

$$E_1 + E_2 = E_1' + E_2' \tag{13.32}$$

where E_1 and E_2 are the electron energies before the collision and E_1' and E_2' are the energies after the collision. The probability of a collision of an electron with energy E_1 and of an electron with energy E_2 is proportional to the probability that there is an electron of energy E_1 and a second electron with energy E_2. If the probability of such a collision is p, then

$$p = B\, f_M(E_1)\, f_M(E_2) \tag{13.33}$$

where B is a constant. The same consideration is valid for particles with energies E_1' and E_2'. Thus, the probability that two electrons with energies E_1' and E_2' collide is given by

$$p' = B\, f_M(E_1')\, f_M(E_2') . \tag{13.34}$$

If the change in energy before and after the collision is ΔE, then $\Delta E = E_1' - E_1$ and $\Delta E = E_2 - E_2'$. Furthermore, if the electron gas is in equilibrium, then $p = p'$ and one obtains

$$f_M(E_1)\, f_M(E_2) = f_M(E_1 + \Delta E)\, f_M(E_2 - \Delta E) . \tag{13.35}$$

Only the exponential function satisfies this condition, that is

$$f_M(E) = A \exp(-\alpha E) \tag{13.36}$$

where α is a positive yet undetermined constant. The exponent is chosen negative to assure that the occupation probability decreases with higher energies. It will become obvious that α is a universal constant and applies to all carrier systems such as electron-, heavy- or light-hole systems.

Next, the constant α will be determined. It will be shown that $\alpha = 1/kT$ using the results of the ideal gas theory. The energy of an electron in an ideal gas is given by

$$E = \tfrac{1}{2} m v^2 = \tfrac{1}{2} m (v_x^2 + v_y^2\, v + v_z^2) . \tag{13.37}$$

The exponential energy distribution of Eq. (13.36) and the normalization condition of Eq. (13.15) yield the normalized velocity distribution

$$f(v_x, v_y, v_z) \quad = \quad \left(\frac{m\,\alpha}{2\,\pi} \right)^{3/2} \exp\left[-\frac{1}{2}\, m\,\alpha\,(v_x^2 + v_y^2 + v_z^2) \right] . \tag{13.38}$$

The average energy of an electron is obtained by (first) calculating the mean-square velocities, $\overline{v_x^2}$, $\overline{v_y^2}$, $\overline{v_z^2}$ from the distribution and (second) using Eq. (13.37) to calculate E from the mean-square velocities. One obtains

$$E \quad = \quad (3/2)\,\alpha^{-1} . \tag{13.39}$$

We now use the result from classic gas theory which states according to Eq. (13.28) that the kinetic energy equals $E = (3/2)\,kT$. Comparison with Eq. (13.39) yields

$$\alpha \quad = \quad (kT)^{-1} \tag{13.40}$$

which concludes the proof of the Maxwell distribution of Eqs. (13.30) and (13.31).

Having determined the value of α, the explicit form of the normalized ***maxwellian velocity distribution*** in cartesian coordinates is

$$\boxed{f_{\mathrm{M}}(v_x, v_y, v_z) \quad = \quad \left(\frac{m}{2\,\pi\,kT} \right)^{3/2} \exp\left[-\frac{\frac{1}{2}\, m\,(v_x^2 + v_y^2 + v_z^2)}{kT} \right]} \tag{13.41}$$

Due to the spherical symmetry of the maxwellian velocity distribution, it is useful to express the distribution in spherical coordinates. For the coordinate transformation we note that $f_{\mathrm{M}}(v_x, v_y, v_z)\,\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z = f_{\mathrm{M}}(v)\,\mathrm{d}v$, and that a volume element $\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z$ is given by $4\,\pi\,v^2\,\mathrm{d}v$ in spherical coordinates. The maxwellian velocity distribution in spherical coordinates is then given by

$$f_{\mathrm{M}}(v) \quad = \quad \left(\frac{m}{2\,\pi\,kT} \right)^{3/2} (4\,\pi\,v^2) \exp\left(-\frac{\frac{1}{2}\, m\,v^2}{kT} \right) . \tag{13.42}$$

The maxwellian velocity distribution is shown in ***Fig***. 13.2. The peak of the distribution, that is the most likely velocity, is $v_{\mathrm{p}} = (2kT/m)^{1/2}$. The mean velocity is given by $\overline{v} = (8kT)/(\pi\,m)^{1/2}$. The root-mean-square velocity can only be obtained by numerical integration.

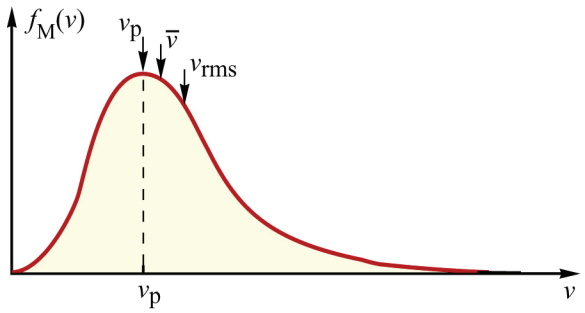


Fig. 13.2. Schematic maxwellian velocity distribution $f_{\mathrm{M}}(v)$ of an ideal elecron gas. The velocity with the highest probability, v_{p}, is lower than the mean velocity, $\overline{v}$, and the root-mean-square velocity, v_{rms}.

13.4 The Boltzmann factor

The maxwellian velocity distribution can be changed to an energy distribution by using the substitution $E = (1/2)\, m\, v^2$. Noting that the energy interval and the velocity interval are related by $\mathrm{d}E = m\, v\, \mathrm{d}v$ and that the number of electrons in the velocity interval, $f_{\mathrm{M}}(v)\, \mathrm{d}v$, is the same as the number of electrons in the energy interval, $f_{\mathrm{MB}}(E)\, \mathrm{d}E$, then the energy distribution is given by

$$f_{\mathrm{MB}}(E) = \frac{2}{\sqrt{\pi}} \frac{\sqrt{E}}{(kT)^{3/2}} e^{-E/kT} \tag{13.43}$$

which is the ***Maxwell–Boltzmann distribution***.

For large energies, the exponential term in the Maxwell–Boltzmann distribution essentially determines the energy dependence. Therefore, the high-energy approximation of the Maxwell–Boltzmann distribution is

$$f_{\mathrm{B}}(E) = A\, e^{-E/kT} \tag{13.44}$$

which is the ***Boltzmann distribution***. The exponential factor of the distribution, $\exp(-E/kT)$, is called the ***Boltzmann factor*** or ***Boltzmann tail***. The Boltzmann distribution does not take into account the quantum mechanical properties of an electron gas. The applicability of the distribution is therefore limited to the classical regime, *i. e.* for $E >> kT$.

13.5 The Fermi–Dirac distribution

In contrast to classical Boltzmann statistics, the quantum mechanical characteristics of an electron gas are taken into account in Fermi–Dirac statistics. The quantum properties which are explicitly taken into account are

- The *wave character* of electrons. Due to the wave character of electrons the Schrödinger equation has only a *finite number of solutions* in the energy interval E and $E + \mathrm{d}E$.

- The *Pauli principle* which states that an eigenstate can be occupied by only two electrons of opposite spin.

Since the Pauli principle strongly restricts the number of carriers per energy level, higher states are populated even at zero temperature. This situation is illustrated in ***Fig***. 13.3, where two electron distributions are illustrated at zero temperature. The distribution in ***Fig***. 13.3(a) does not take into account the Pauli principle while that in ***Fig***. 13.3(b) does.

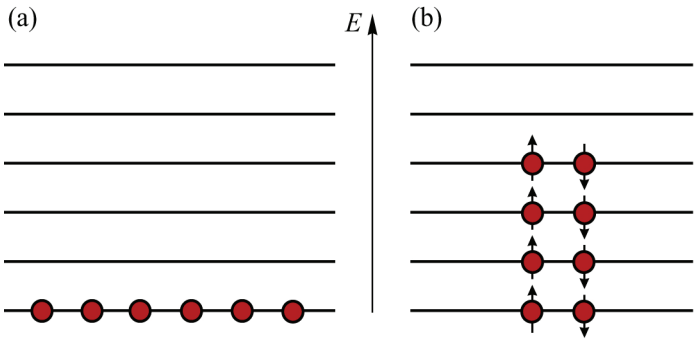


Fig. 13.3. Distribution of electrons at zero temperature among discrete energy levels (a) without Pauli principle and (b) with Pauli principle and spin taken into account. Spin 'up' and 'down' is illustrated by arrows.

The first restriction imposed by quantum mechanics is the *finiteness of states* within an energy interval E and $E + dE$. Recall that the finiteness of states played a role in the derivation of the density of states. The density of states in an isotropic semiconductor was shown to be

$$\rho_{DOS}(E) = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E} \tag{13.45}$$

where E is the kinetic energy. Note that for the derivation of the density of states the Pauli principle has been taken into account. Therefore, the states given by Eq. (13.45) can be occupied only by *one* electron. Since the number of states per velocity-interval will be of interest, Eq. (13.45) is modified using $E = (½)\, mv^2$ and $dE = mv\, dv$. Note that the number of states per energy interval dE is the same as the number of states per velocity interval dv, *i. e.* $\rho_{DOS}(E)\, dE = \rho_{DOS}(v)\, dv$. The number of states per velocity interval (and per unit volume) is then given by

$$\rho_{DOS}(v) = \frac{m^{*3}}{\pi^2\,\hbar^3}v^2 \tag{13.46}$$

for an isotropic semiconductor.

The *Fermi–Dirac distribution*, also called the Fermi distribution, gives the probability that a state of energy E is occupied. Since the Pauli principle has been taken into account in the density of states given by Eq. (13.45), each state can be occupied by at most one electron. The Fermi distribution is given by

$$f_F(E) = \left[1 + \exp\left(\frac{E - E_F}{kT}\right)\right]^{-1} \tag{13.47}$$

where E_F is called the Fermi energy. At $E = E_F$ the Fermi distribution has a value of 1/2. For small energies the Fermi distribution approaches 1; thus low-energy states are very likely to be populated by electrons. For high energies the Fermi distribution decreases exponentially; states of high energy are less likely to the populated. Particles which follow a Fermi distribution are called ***fermions***. Electrons and holes in semiconductors are such fermions. A system of particles which obey ***Fermi*** statistics are called a ***Fermi gas***. Electrons and holes constitute such Fermi gases.

An approximate formula for the Fermi distribution can be obtained for high energies. One obtains for $E >> E_F$

$$f_F(E) \approx \exp\left(-\frac{E - E_F}{kT}\right) = f_B(E)\ . \tag{13.48}$$

This distribution coincides with the Boltzmann distribution. Thus the (quantum-mechanical) Fermi distribution and the (classical) Boltzmann distribution coincide for high energies, *i. e.* in the classical regime.

Next we prove the Fermi distribution of Eq. (13.47) by considering a collision between two electrons. For simplification we assume that one of the electrons has such a high energy that it belongs to the classical regime of semiconductor statistics. Quantum statistics applies to the other low-energy electron. During the collision of the two electrons, the energy is conserved

$$E_1 + E_2 = E_1' + E_2' \tag{13.49}$$

where, as before (Eq. 13.32), E_1 and E_2 are electron energies before the collision and E_1' and E_2' are the energies after the collision.

The probability for the transition $(E_1, E_2) \rightarrow (E_1', E_2')$ is given by

$$p = f_F(E_1)\ f_B(E_2)\ \left[1 - f_F(E_1')\right]\left[1 - f_B(E_2')\right] \tag{13.50}$$

where it is assumed that E_2 and E_2' are relatively large energies and the corresponding electron can be properly described by the Boltzmann distribution. The terms $[\,1 - f_F(E_1')\,]$ and $[\,1 - f_B(E_2')\,]$ describe the probability that the states of energies E_1' and E_2' are empty, and are available for the electron after the collision. Further simplification is obtained by considering that E_2' is large and therefore $[1 - f_B(E_2')] \approx 1$. Equation (13.50) then simplifies to

$$p = f_F(E_1)\, f_B(E_2)\left[1 - f_F(E_1')\right]. \tag{13.51}$$

The same considerations are valid for the transition $(E_1', E_2') \rightarrow (E_1, E_2)$. The probability of this transition is given by

$$p' = f_F(E_1')\ f_B(E_2')\ \left[1 - f_F(E_1)\right]. \tag{13.52}$$

Under equilibrium conditions both transition probabilities are the same, *i. e.* $p = p'$. Equating Eqs. (13.51) and (13.52), inserting the Boltzmann distribution for $f_B(E)$, and dividing by $f_F(E_1)$ $f_F(E_1')\, f_B(E_2)$ yields

$$\frac{1}{f_F(E_1')} - 1 = \left[\frac{1}{f_F(E_1)} - 1\right] \exp\left[\frac{E_2 - E_2'}{kT}\right] \tag{13.53}$$

which must hold for all E_1 and E_1'. This condition requires that

$$\frac{1}{f_F(E)} - 1 = A \exp\frac{E}{kT} \tag{13.54}$$

where A is a constant. If the value of the constant is taken to be $A = \exp(-E_F / kT)$ one obtains the ***Fermi–Dirac distribution***

$$\boxed{f_F(E) = \left[1 + \exp\left(\frac{E - E_F}{kT}\right)\right]^{-1}} \tag{13.54}$$

which proves Eq. (13.47).

The Fermi–Dirac distribution is shown for different temperatures in ***Fig***. 13.4. At the energy $E = E_F$ the probability of a state being populated has always a value of ½ independent of temperature. At higher temperatures, states of higher energies become populated. Note that the Fermi–Dirac distribution is symmetric with respect to E_F, that is

$$f_F(E_F + \Delta E) = 1 - f_F(E_F - \Delta E) \tag{13.55}$$

where ΔE is any energy measured with respect to the Fermi energy.

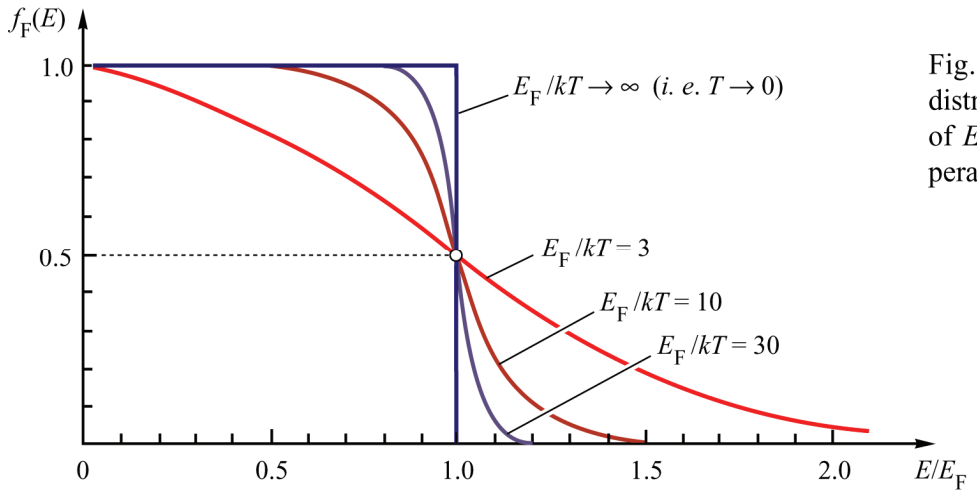


Fig. 13.4. Fermi–Dirac distribution as a function of E/E_F for different temperatures.

The Fermi–Dirac velocity distribution of the particles in a Fermi gas is obtained by multiplication of Eq. (13.46) with Eq. (13.47)

$$g(v) \;=\; \rho_{DOS}(v)\; f_F(v) \;=\; \frac{m^3\, v^3}{\pi^2\, \hbar^3}\left[1 \,+\, \exp\left(\frac{\frac{1}{2}\, m\, v^2 - E_F}{kT}\right)\right]^{-1} \tag{13.56}$$

where we have used the fact that the energy of the Fermi gas is purely kinetic, *i. e.* $E = (1/2)\, mv^2$. Note that $g(v)$ is the number of carriers per velocity interval v and $v + \mathrm{d}v$ and per unit volume. If the velocity v is expressed in terms of its components, then the spherical volume element, $4\pi\, v^2\, \mathrm{d}v$, is modified to a volume element in rectangular coordinates, $\mathrm{d}v_x\, \mathrm{d}v_y\, \mathrm{d}v_z$. Thus, using $g(v)\, \mathrm{d}v = g(v_x, v_y, v_z)\, \mathrm{d}v_x\, \mathrm{d}v_y\, \mathrm{d}v_z$, one obtains

$$g(v_x, v_y, v_z) \;=\; \frac{m^3}{\pi^2\, \hbar^3}\; \frac{1}{4\,\pi}\, v \left\{1 \,+\, \exp\left[\frac{\frac{1}{2}\left(v_x^2 + v_y^2 + v_z^2\right) \,-\, E_F}{kT}\right]\right\}^{-1} \tag{13.57}$$

which is the Fermi velocity distribution (per unit volume) in cartesian coordinates.

The Fermi distribution of energies of an ideal gas is obtained by multiplication of Eq. (13.45) with Eq. (13.47) and is given by

$$g(E) \;=\; \frac{1}{2\,\pi^2}\left(2\,\frac{m}{\hbar^2}\right)^{3/2} \sqrt{E}\left[1 \,+\, \exp\left(\frac{E - E_F}{kT}\right)\right]^{-1} \tag{13.59}$$

when $g(E)$ is the number of particles in the energy interval E and $E + \mathrm{d}E$ and per unit volume.

13.6 The Fermi–Dirac integral of order $j = +1/2$ (3D semiconductors)

The Fermi–Dirac integral of order $j = +1/2$ allows one to calculate the free carrier concentration in a three-dimensional (3D) semiconductor. The free carrier concentration in one band, *e. g.* the conduction band, of a semiconductor is obtained from the product of density of states and the

state occupation probability, *i. e.*

$$n = \int_{E_C}^{E_{top}} \rho_{DOS}(E)\, f_F(E)\, dE \ . \tag{13.60}$$

Integration over all conduction band states is required to obtain the total concentration. The upper limit of integration is the top of the conduction band and can be extended to infinity. This extension of $E_C^{top} \to \infty$ can be done without losing accuracy, since $f_F(E)$ converges strongly at high energies. The two functions, $\rho_{DOS}(E)$, $f_F(E)$, and their product are schematically shown in ***Fig***. 13.5 for a semiconductor with three, two, and one, spatial degrees of freedom. The concentration per unit energy $n(E)$ is the product of state density and distribution function.

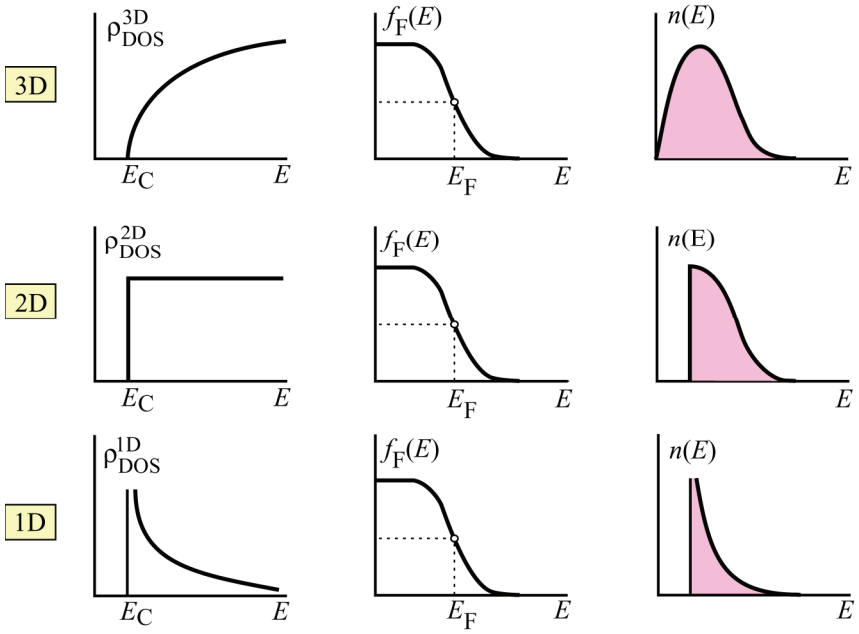


Fig. 13.5. Density of states (ρ_{DOS}), Fermi-Dirac distribution function (f_F) and carrier concentration (n) as a function of energy for a 3D, 2D, and 1D system. The shaded areas represent the total carrier concentration in the conduction band.

Equation (13.60) is evaluated by inserting the explicit expressions for the state density and the Fermi–Dirac distribution (Eq. 13.47). One obtains

$$n = \frac{1}{2\,\pi^2} \left(\frac{2\,m^*\,kT}{\hbar^2} \right)^{3/2} \int_0^{\infty} \frac{\eta^{1/2}}{1 + \exp(\eta - \eta_F)}\, d\eta \tag{13.61}$$

where $\eta = E / kT$ and $\eta_F = -(E_C - E_F) / kT$ is the reduced Fermi energy. η_F is positive when E_F is inside the conduction band. The equation can be written in a more convenient way by using the Fermi–Dirac integral of order j, which is defined by (Sommerfeld, 1928; Sommerfeld and Frank, 1931)

$$F_j(\eta_F) = \frac{1}{\Gamma(j+1)} \int_0^{\infty} \frac{\eta^j}{1 + \exp(\eta - \eta_F)}\, d\eta \tag{13.62}$$

where $\Gamma(j+1)$ is the Gamma-function.

Sommerfeld's original definition of the Fermi–Dirac integral omitted the term of the Gamma – function, *i. e.* $F_j^s(\eta_F) = \int_0^\infty \eta^j / [1 + \exp(\eta - \eta_F)]\, d\eta$. The modern definition of the Fermi–Dirac integral of Eq. (13.62) has the following advantages: (*i*) Unlike F_j^s, the functions F_j exist for negative orders of *j*, *e. g.* $j = -(1/2)$, -1, $-3/2$ etc. (*ii*) In the non-degenerate limit in which $\eta_F << 0$, all members of the $F_j(\eta_F)$ family reduce to $F_j(\eta_F) \rightarrow \exp \eta_F$ for all *j*. (*iii*) The *derivative* of the Fermi–Dirac integral of integer order *j* can be expressed as a Fermi–Dirac integral of order $(j - 1)$, *i. e.* $(\partial / \partial\eta_F)\, F_j(\eta_F) = F_{j-1}(\eta_F)$.

With $j = (1/2)$ and $\Gamma(3/2) = \pi^{1/2} / 2$ one obtains

$$n = N_c\, F_{1/2}(\eta_F) \tag{13.63}$$

where N_c is the effective state density at the bottom of the conduction band. The Fermi–Dirac integral $F_{1/2}$ is shown in ***Fig***. 13.6 along with several approximations which will be discussed later.

For $j = (1/2)$ the Fermi–Dirac integral is

$$F_{1/2}(\eta_F) = \frac{1}{\Gamma(3/2)} \int_0^\infty \frac{\eta^{1/2}}{1 + \exp(\eta - \eta_F)}\, d\eta \ . \tag{13.64}$$

The evaluation of the integral cannot be done analytically. Even though the numerical calculation of the Fermi–Dirac integral is straightforward, it proves frequently convenient to use *approximate* analytic solutions of $F_{1/2}(\eta_F)$.

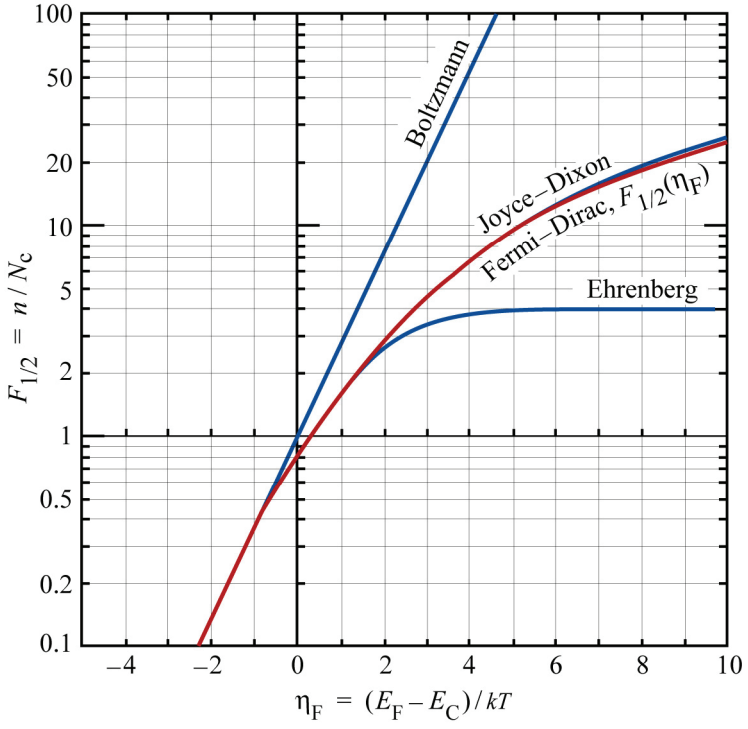


Fig. 13.6. Fermi–Dirac integral of order 1/2 as a function of the reduced Fermi energy η_F. Also shown are the Boltzmann distribution, the Joyce–Dixon approximation, and the Ehrenberg approximation.

For analytic approximations of the Fermi–Dirac integral $n / N_c = F_{1/2}(\eta_F)$, the inverse function is frequently used, that is the reduced Fermi energy η_F is expressed as a function of n / N_c. A

number of analytic approximations developed prior to 1982 have been reviewed by Blakemore (Blakemore, 1982). To classify various approximations, we differentiate between non-degeneracy and degeneracy. In the *non-degenerate* regime, the Fermi energy is below the bottom of the conduction band, $E_F << E_C$. In the *degenerate* regime the Fermi energy is at or above the bottom of the conduction band.

- ***Extreme non-degeneracy (3D)***
  In the case of extreme non-degeneracy (*i. e.* $E_C - E_F >> kT$ or $F_{1/2} << 1$ or $n << N_c$) the Fermi–Dirac distribution approaches the Boltzmann distribution. One obtains

$$\eta_F = -\frac{E_C - E_F}{kT} \approx \ln\frac{n}{N_c} \tag{13.65}$$

  which is shown in ***Fig***. 13.7. This approximation is good when the Fermi energy is 2 *kT* or more below the bottom of the conduction band. Rearrangement of the equation yields the carrier concentration as a function of the Fermi energy in the non-degenerate limit

$$n = N_c \exp\left(-\frac{E_C - E_F}{kT}\right) \tag{13.66}$$

- ***Extreme degeneracy (3D)***
  In the case of extreme degeneracy (*i. e.* $(E_F - E_C) >> kT$ or $F_{1/2} >> 1$ or $n >> N_c$) the Fermi–Dirac integral reduces to

$$\eta_F = -\frac{E_C - E_F}{kT} \approx \left[\frac{3}{2}\Gamma\left(\tfrac{3}{2}\right)\frac{n}{N_c}\right]^{2/3} \approx \left(\frac{3}{4}\sqrt{\pi}\,\frac{n}{N_c}\right)^{2/3} \tag{13.67}$$

  which is shown in ***Fig***. 13.7. The range of validity for this approximation is $E_F - E_C > 10\ kT$, *i. e.* when the Fermi energy is well within the conduction band. Rearrangement of the equation and using the effective density of states (N_c) yields the carrier concentration as a function of the Fermi energy in the degenerate limit

$$n = \frac{1}{3\pi^2}\left(\frac{2\,m^*(E_F - E_C)}{\hbar^2}\right)^{3/2} \tag{13.68}$$

- ***The Ehrenberg Approximation (3D)***
  This approximation (Ehrenberg, 1950) was developed for weak degeneracy and is shown in ***Fig***. 13.6. The approximation is given by

$$\eta_F = -\frac{E_C - E_F}{kT} \approx \ln\frac{n}{N_c} - \ln\left(1 - \frac{1}{4}\,\frac{n}{N_c}\right). \tag{13.69}$$

  For small *n* the second logarithm term approaches zero, that is the Boltzmann distribution is recovered. The range of validity of the approximation is limited to $E_F - E_C \leq 2\ kT$, *i. e.* to weak degeneracy.

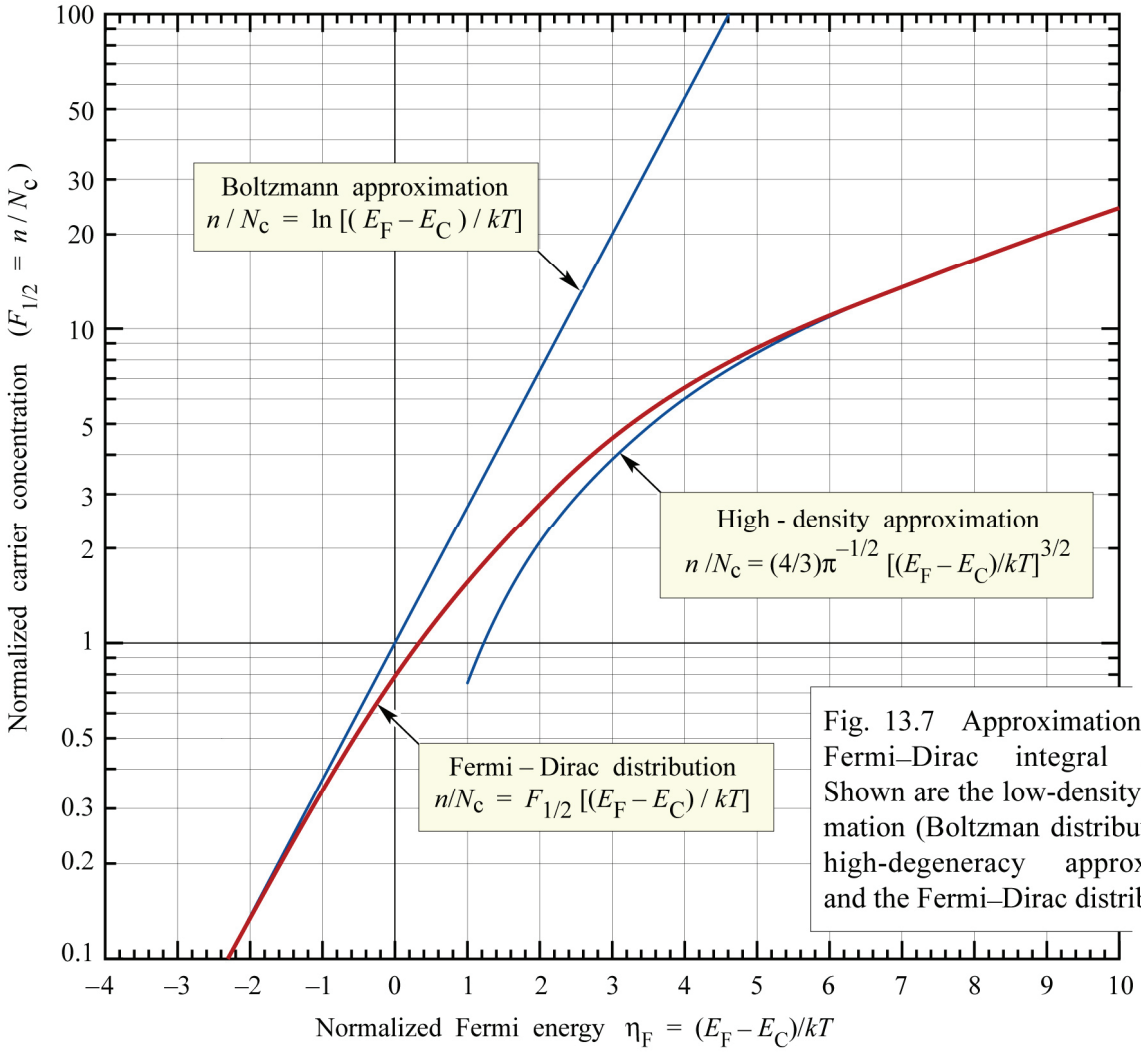


Fig. 13.7 Approximations for the Fermi–Dirac integral $F_{1/2}(\eta_F)$. Shown are the low-density approximation (Boltzman distribution), the high-degeneracy approximation, and the Fermi–Dirac distribution.

- ***The Joyce – Dixon approximation (3D)***

An approximation valid for a wider range of degeneracy was developed by Joyce and Dixon (Joyce and Dixon, 1977; Joyce, 1978). This approximation expresses the reduced Fermi energy as a sum of the Boltzmann term and a polynomial, *i. e.*

$$\eta_F = -\frac{E_C - E_F}{kT} \approx \ln \frac{n}{N_c} + \sum_{m=1}^{4} A_m \left(\frac{n}{N_c} \right)^m \tag{13.70}$$

where the first four coefficients A_m are given by

$$\begin{aligned} A_1 &= \sqrt{2}\,/\,4 = 3.53553 \times 10^{-1}, \\ A_2 &= -4.95009 \times 10^{-3}, \\ A_3 &= 1.48386 \times 10^{-4}, \\ A_4 &= -4.42563 \times 10^{-6}. \end{aligned} \tag{13.71}$$

The Joyce – Dixon approximation given here is shown in ***Fig***. 13.6 and can be used for degeneracies of $E_F - E_C \approx 8\ kT$. Inclusion of higher terms in the power series ($m > 4$) allows one to extend the Joyce – Dixon approximation to higher degrees of degeneracy.

- ***The Chang–Izabelle Approximation (3D)***
The Chang–Izabelle approximation (Chang and Izabelle, 1989) is a full-range approximation which is valid for non-degenerate as well as degenerate semiconductors. The approximation is motivated by the fact that low-density and high-density approximations are available (see Eqs. 13.66 and 13.68) which are the exact solutions in the two extremes. The Chang–Izabelle approximation represents the construction of a function, which approaches the low-density solution and the high-density solution of the Fermi–Dirac integral as shown in ***Fig***. 13.7. The reduced Fermi energy is then given by

$$\eta_F = \frac{E_C - E_F}{kT} \approx \ln \frac{n}{N_c + n} \left[\frac{3}{2}\ \Gamma\left(\tfrac{3}{2}\right) \right]^{2/3} \frac{n/N_c}{(A + n/N_c)^{1/3}} \tag{13.72}$$

where

$$n/N_c = F_{1/2}\ , \tag{13.73}$$

$$\frac{3}{2}\ \Gamma\left(\tfrac{3}{2}\right) = \frac{3}{2}\ \frac{\sqrt{\pi}}{2} = 1.32934\ , \tag{13.74}$$

$$A = \frac{\left[(3/2)\ \Gamma\ (3/2)\right]^2\ (n_0/N_c)^3}{\left[\ln\ (1\ +\ N_c/n_0)\right]^3} - \frac{n_0}{N_c}\ , \tag{13.75}$$

$$\frac{n_0}{N_c} = \frac{n(E_F = E_C)}{N_c} = F_{1/2}(\eta_F = 0) = 0.76515\ . \tag{13.76}$$

One can easily verify that Eq. (13.72) recovers the low-density approximation and the high-density approximation for $n << N_c$ and $n >> N_c$, respectively. Furthermore, the approximation yields an exact solution for $\eta_F = 0$, *i. e.* when the Fermi energy touches the bottom of the conduction band. The largest relative error of η_F is 1 % in the Chang–Izabelle approximation. Chang and Izabelle (1989) showed that the relative error can be further reduced by a weighting function and a polynomial function. Using these functions, the maximum relative error is reduced to 0.033%.

- ***The Nilsson approximation (3D)***
The Nilsson approximation (Nilsson, 1973) is valid for the entire range of Fermi energies. It is given by

$$\eta_F = -\frac{\ln\ (n/N_c)}{n/N_c - 1} + \left(\frac{3}{4}\sqrt{\pi}\ \frac{n}{N_c} \right)^{2/3} + \frac{(3/2)\sqrt{\pi}\ (n/N_c)}{\left[3 + (3/4)\sqrt{\pi}\ (n/N_c)\right]^2}\ . \tag{13.77}$$

The maximum relative error of the approximation is 1.1%.

13.7 The Fermi–Dirac integral of order $j = 0$ (2D semiconductors)

The Fermi–Dirac integral of order $j = 0$ allows one to calculate the free carrier density in a two-dimensional (2D) semiconductor. For semiconductor structures with only two degrees of spatial freedom, the Fermi–Dirac integral is obtained from Eq. (13.60) by insertion of the two-dimensional density of states. One obtains for the 2D carrier density

$$n^{2D} = \int_{E_C}^{E_{top}} \rho_{DOS}^{2D}(E)\ f_F(E)\ dE = \frac{m^* kT}{\pi \hbar^2} \int_0^{\infty} \left[1 + \exp(\eta - \eta_F)\right]^{-1}\ d\eta \qquad (13.78)$$

where $\eta = E / kT$ and $\eta_F = (E_F - E_C) / kT$ are reduced energies. The integral can be written as the Fermi–Dirac integral of zero ($j = 0$) order

$$F_{j=0}(\eta_F) = \frac{1}{\Gamma(1)} \int_0^{\infty} \left[1 + \exp(\eta - \eta_F)\right]^{-1}\ d\eta \qquad (13.79)$$

where $\Gamma(1) = 1$ is the Gamma-function. The two-dimensional carrier density can be written by using the effective density of states of a 2D system (N_c^{2D}). One obtains

$$n^{2D} = N_c^{2D}\ F_0(\eta_F) \qquad (13.80)$$

which is formally similar to the corresponding equation in three dimensions (Eq. 13.63). The Fermi–Dirac integral of zero order ($j = 0$) can be solved analytically. Using the integral formula

$$\int \frac{dx}{1 + e^x} = -\ln(1 + e^{-x})\ , \qquad (13.81)$$

one obtains

$$F_0(\eta_F) = \ln(1 + e^{\eta_F})\ . \qquad (13.82)$$

Thus, the two-dimensional carrier density depends on the reduced Fermi energy according to

$$\frac{n^{2D}}{N_c^{2D}} = \ln(1 + e^{\eta_F})\ . \qquad (13.83)$$

Rearrangement of the equation yields the Fermi energy as a function of the carrier density

$$\eta_F = -\frac{E_C - E_F}{kT} = \ln\left[\exp\left(\frac{n^{2D}}{N_c^{2D}}\right) - 1\right] \qquad (13.84)$$

- ***Extreme non-degeneracy (2D)***
  Approximation for the low-density regime ($n^{2D} << N_c^{2D}$) and the high-density regime ($n^{2D} >> N_c^{2D}$) can be easily obtained from Eqs. (13.83) and (13.84). In the low-density regime one obtains

$$\eta_F = -\frac{E_C - E_F}{kT} = \ln \frac{n^{2D}}{N_c^{2D}} \tag{13.85}$$

which is valid if the Fermi energy is much below the bottom of the conduction subband. Rearrangement of the equation yields the two-dimensional carrier density as a function of the Fermi energy in the non-degenerate limit

$$n^{2D} = N_c^{2D} \exp\left(-\frac{E_C - E_F}{kT}\right). \tag{13.86}$$

- ***Extreme degeneracy (2D)***
  In the high-density regime one obtains

$$\eta_F = -\frac{E_C - E_F}{kT} = \frac{n^{2D}}{N_c^{2D}} \tag{13.87}$$

Rearrangement of the equation and insertion of the explicit expression for N_c^{2D} yields

$$n^{2D} = \frac{m^*}{\pi \hbar^2} \left(E_F - E_C\right) \tag{13.88}$$

that is the Fermi energy and the two-dimensional density follow a linear relation for two-dimensional structures.

13.8 The Fermi–Dirac integral of order $j = -1/2$ (1D semiconductors)

The Fermi–Dirac integral of order $j = -1/2$ allows one to calculate the free carrier density (per unit length) in a one-dimensional (1D) semiconductor. In semiconductor structures with only one degree of spatial freedom, the Fermi–Dirac integral is obtained from Eq. (13.60) by insertion of the 1D density of states. One obtains

$$n^{1D} = \int_{E_C}^{E_{top}} \rho_{DOS}^{1D}(E)\, f_F(E)\, dE = \frac{kT}{\pi \hbar} \sqrt{\frac{m^*}{2\,kT}} \int_0^\infty \frac{\eta^{-1/2}}{1 + \exp(\eta - \eta_F)}\, d\eta \tag{13.89}$$

where $\eta = E/kT$ and $\eta_F = -(E_C - E_F)/kT$ are reduced energies. The integral can be written as the Fermi–Dirac integral of order j = – 1/2

$$F_{-1/2}(\eta_F) = \frac{1}{\Gamma(1/2)} \int_0^\infty \frac{\eta^{-1/2}}{1 + \exp(\eta - \eta_F)} \tag{13.90}$$

where $\Gamma(1/2) = \pi^{1/2}$ is the Gamma-function. Using the effective density of states of a one-dimensional system (N_c^{1D}), the one-dimensional density can be written as

$$n^{1D} = N_c^{1D}\, F_{-1/2}(\eta_F) \tag{13.91}$$

which is similar to the corresponding equations in three (Eq. 13.63) and two (Eq. 13.80) dimensions. The Fermi–Dirac integral of order $j = -1/2$ can only be obtained by numerical integration or by approximate solutions which will be discussed in the following sections. The $j = -1/2$ Fermi–Dirac integral has asymptotic solutions for the regimes of non-degeneracy and high degeneracy.

- ***Extreme non-degeneracy (1D)***
  In the regime of extreme non-degeneracy ($n^{1D} << N_c^{1D}$) the Fermi–Dirac integral of order $j = -1/2$ approaches the Boltzmann distribution. One obtains

$$\eta_F = -\frac{E_C - E_F}{kT} = \ln \frac{n^{1D}}{N_c^{1D}} \tag{13.92}$$

  This approximation is good for $E_F - E_C \leq 2\ kT$, *i. e.* when the Fermi energy is at least $2\,kT$ below the bottom of the conduction subband. Rearrangement of the equation yields the one-dimensional carrier density as a function of the Fermi energy in the non-degenerate limit

$$n^{1D} = N_c^{1D} \exp\left(-\frac{E_C - E_F}{kT}\right). \tag{13.93}$$

- ***Extreme degeneracy (1D)***
  In the case of extreme degeneracy ($n^{1D} >> N_c^{1D}$) the Fermi–Dirac integral of order $j = -1/2$ reduces to

$$\eta_F = -\frac{E_C - E_F}{kT} \approx \frac{1}{4}\left[\Gamma\left(\frac{1}{2}\right)\right]^2 \left(\frac{n^{1D}}{N_c^{1D}}\right)^2 \approx \frac{\pi}{4}\left(\frac{n^{1D}}{N_c^{1D}}\right)^2 \tag{13.94}$$

  The range of validity of the approximation is $E_F - E_C > 10\ kT$, *i. e.* when the Fermi energy is well within the conduction band.

Exercise 1: ***The Fermi-Dirac distribution***. Show that the Fermi-Dirac Distribution is an odd-symmetry function with respect to the point $E = E_F$, i.e. show that the following equation is correct: $f_F(E_F + \Delta E) = 1 - f_F(E_F - \Delta E)$

Solution: Writing $f_F(E_F + \Delta E) = 1 - f_F(E_F - \Delta E)$ by using the Fermi–Dirac distribution, we obtain

$$\left[1 + \exp\left(\frac{(E_F + \Delta E) - E_F}{kT}\right)\right]^{-1} = 1 - \left[1 + \exp\left(\frac{(E_F - \Delta E) - E_F}{kT}\right)\right]^{-1}$$

Simplification yields

$$\left[1 + \exp\left(\frac{\Delta E}{kT}\right)\right]^{-1} = 1 - \left[1 + \exp\left(\frac{-\Delta E}{kT}\right)\right]^{-1}$$

$$\left[1 + \exp\left(\frac{\Delta E}{kT}\right)\right]^{-1} = \frac{1 + \exp\left(\frac{-\Delta E}{kT}\right) - 1}{1 + \exp\left(\frac{-\Delta E}{kT}\right)}$$

$$\left[1 + \exp\left(\frac{\Delta E}{kT}\right)\right]^{-1} = \frac{\exp\left(\frac{-\Delta E}{kT}\right)}{1 + \exp\left(\frac{-\Delta E}{kT}\right)}$$

$$\left[1 + \exp\left(\frac{\Delta E}{kT}\right)\right]^{-1} = \frac{1}{\exp\left(\frac{\Delta E}{kT}\right) + \exp\left(\frac{\Delta E}{kT}\right)\exp\left(\frac{-\Delta E}{kT}\right)}$$

$$\left[1 + \exp\left(\frac{\Delta E}{kT}\right)\right]^{-1} = \left[1 + \exp\left(\frac{\Delta E}{kT}\right)\right]^{-1}.$$

Because both sides of the equation are identical, the equation is correct, and we have shown what was to be shown.

Exercise 1: ***Approximate rules***. Which or the following statements or approximate rules are correct?

(a) In an n-type semiconductor, for $E_C > E_F$, we can use the Boltzmann distribution for the state occupancy in the conduction band.

(b) In an n-type semiconductor, for $E_F > E_C$, we must use the Fermi–Dirac distribution for the state occupancy in the conduction band.

(c) In an intrinsic semiconductor, the Fermi level is *approximately* in the middle of the forbidden gap.

(d) What is the physical reason that the Fermi level is *approximately* but *not exactly* in the middle of the forbidden gap?

Solution: (a)–(c) are correct. (d) Because the density of states is different in the conduction and valence band, and because the Fermi distribution is symmetric with respect to the point $E = E_F$, the Fermi level is *not exactly* in the middle of the forbidden gap.

References

Blakemore J. S. "Approximations for Fermi-Dirac integrals, especially the function $F_{1/2}(\eta)$ used to describe electron density in a semiconductor" *Sol. State Electronics* **25,** 1067 (1982)

Chang T. Y. and Izabelle A. "Full range analytic approximations for Fermi energy and Fermi–Dirac integral $F_{-1/2}$ in terms of $F_{1/2}$" *Journal of Applied Physics* **65**, 2162 (1989)

Ehrenberg, W. "The electrical conductivity of simple semiconductors" *Proceedings of the Physical Society of London, A***63**, 75 (1950)

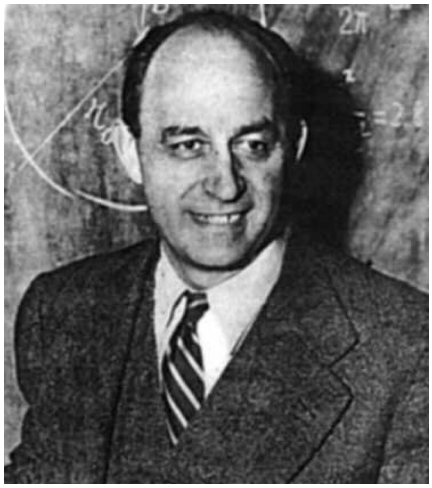
Joyce W. B. and Dixon R. W. "*Analytic approximations for the Fermi energy of an ideal Fermi gas" Applied Physics Letters* **31**, 354 (1977)

Joyce W. B. "Analytic approximations for the Fermi energy in AlGaAs" *Applied Physics Letters* **32**, 680 (1978)

Kittel C. and Kroemer H. *Thermal Physics*, 2nd edition (Freeman, San Francisco, 1980)
Nilsson N. G. "An accurate approximation of the generalized Einstein relation for degenerate semiconductors" *Physica Status Solidi A***19**, K75 (1973)
Sommerfeld A. "Zur Elektronentheorie der Metalle aufgrund der Fermischen Statistik; Teil I: Allgemeines, Stromungs- und Austrittsvorgange" (translated title: "On the electron theory in metal – based on Fermi's statistics; Part 1: General, current and emission processes") *Zeitschrift Physik* **47**, 1 (1928)
Sommerfeld A. and Frank N. H. "The statistical theory of thermoelectric, galvano- and thermo-magnetic phenomena in metals" *Review of Modern Physics* **3**, 1 (1931)

James Maxwell (1831–1879)
Established velocity distribution of gases

Ludwig Boltzmann (1844–1906)
Established classical statistics



Enrico Fermi (1901–1954)
Established quantum statistics

14

Carrier concentrations

The *activation energy* of impurities will be frequently used in this chapter. It is useful to recall the interdependence of free energy, internal energy, enthalpy, entropy, and activation energy. To do so, consider the electronic ionization of an impurity, for example a donor

$$D^0 \quad \leftrightarrow \quad D^+ + \mathrm{e} \; . \tag{14.1}$$

The effective work necessary to accomplish the ionization process at a constant temperature and pressure equals to the change of ***Gibbs free energy*** of the system (Kittel and Kroemer, 1980; Reif, 1965). In thermodynamics, Gibbs free energy G is defined as

$$G \;\; = \;\; H \; - \; T \, S \; , \tag{14.2}$$

$$H \;\; = \;\; E \; + \; P \, V \tag{14.3}$$

where H is the reaction enthalpy, S the entropy, E the internal energy, and PV the product of pressure and volume of the system. The change in Gibbs free energy occurring during the donor ionization process of Eq. (14.1) at a constant temperature T is then given by

$$\Delta G \;\; = \;\; \Delta H \; - \; T \Delta S \; , \tag{14.4}$$

$$\Delta H \;\; = \;\; \Delta E \; + \; P \, \Delta V \tag{14.5}$$

where constant temperature and constant pressure is assumed. Gibbs free energy is the proper energy to be used in a Boltzmann factor or Fermi function (see Sect. on *semiconductor statistics*). The change in volume of the system occurring during chemical reactions can be quite significant. However, the change in volume during the electronic reaction of Eq. (14.1) is very small since the valence electron configuration does not change. The change in volume can therefore be neglected. In this chapter, the change in entropy as well as the mechanical work ($P \, \Delta V$) will be neglected. In this case, it is $\Delta G \approx \Delta H \approx \Delta E$. The energy required for the ionization reaction of Eq. (14.1) is the difference in internal energy, *i. e*. the difference in energy of states occupied by the electron before and after the ionization process. The change in free energy for donors can then be written as $\Delta G \approx \Delta H \approx \Delta E = E_C - E_D = E_d$, that is, the ionization energy equals the donor level energy relative to the bottom of the conduction band. The enthalpy and the entropy of ionization of centers in semiconductors were further considered by Thurmond (1975) and by Van Vechten and Thurmond (1976a, 1976b). The authors made simple estimates of the entropy of ionization of coulombic, isoelectronic, and vacancy-type defects in semiconductors by considering the effect of localized and free-carrier charge distributions upon the lattice modes. The empirical values of these entropies are observed as the temperature variation of the corresponding ionization levels (*i. e.* the term $T \, \Delta S$ in Eq. 14.4). The change in entropy during

the ionization reaction of Au-related levels in Si was considered by Lang *et al*. (1980), who differentiated between the entropy change due to electronic degeneracy and due to atomic vibrational changes. The authors showed that the change in entropy can be a small fraction (10 %) of the ionization enthalpy.

Typical densities of free carriers in semiconductors range from 10^{15} cm^{-3} to 10^{20} cm^{-3}. It is impossible to describe the energies or velocities of those carriers individually. An alternative to the individual characterization of particles is the *statistical* description of a carrier system. The statistical description uses *probabilities* of velocities or energies rather than knowing these quantities for all individual carriers. Thus, the statistical treatment represents a simplification. The derivation of the energy distribution function treats the carrier system as an *ideal gas*, for example a gas of oxygen molecules. The ideal gas is assumed to have only *elastic* collisions between atoms or molecules. Furthermore, the energy of the gas molecules is assumed to be purely translational *kinetic*. Since these properties are applied to the electron or hole system, those systems are frequently referred to as *electron-gases* or *hole-gases*.

The free carrier concentration in semiconductors depends on a number of parameters such as the doping concentration, impurity activation energy, temperature, and other parameters. Given the results of the previous sections on the density of states and the distribution functions, the carrier concentration can now be calculated. In the calculation intrinsic, extrinsic, and compensated semiconductors will be considered.

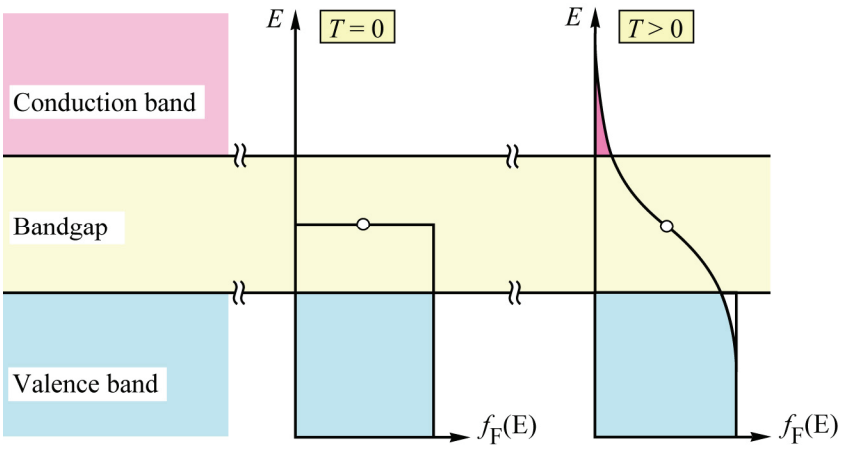


Fig. 14.1. Carrier distributions in the conduction band and valence band and Fermi distribution function in an intrinsic semiconductor at $T = 0$ K and at finite temperatures $T > 0$ K.

14.1 Intrinsic semiconductors

The carrier concentration of pure, undoped semiconductors is determined by thermal excitation of electrons from the valence band to states in the conduction band. An intrinsic semiconductor has a filled valence band and an empty conduction band at zero temperature. This property is the very definition of semiconductors. The band diagram along with the $T = 0$ K Fermi distribution function is shown in ***Fig***. 14.1.

As the temperature increases, a small fraction of electrons in the valence band is excited into the conduction band. Thus the number of holes (unoccupied states) p in the valence band coincides with the number of electrons n in the conduction band. Semiconductors for which $n = p$ are called *intrinsic*. The condition that the concentration of electrons coincides with the concentration of holes requires that the Fermi energy be within the forbidden gap. The position of the Fermi energy in the gap is visualized in ***Fig***. 14.1. The electron and hole concentrations are given by

$$n = p \,, \tag{14.6}$$

$$N_c \, F_{1/2}(\eta_F) = N_v \, F_{1/2}(\eta_F) \,. \tag{14.7}$$

Since the Fermi energy is within the forbidden gap, *i. e.* many values of kT below the conduction band and many values of kT above the valence band, simpler Boltzmann statistics can be used instead of Fermi–Dirac statistics. Equation (14.7) then simplifies to

$$N_c \exp\left(-\frac{E_C - E_F}{kT}\right) = N_v \exp\left(-\frac{E_F - E_V}{kT}\right) . \tag{14.8}$$

Rearrangement of the equation and the definition of the gap energy $E_g = E_C - E_V$ yields for the Fermi energy of an *intrinsic* semiconductor

$$E_F = E_V + \frac{1}{2} E_g + \frac{kT}{2} \ln \frac{N_v}{N_c} \,. \tag{14.9}$$

In this equation, $E_V + (1/2) E_g$ represents the mid-gap energy. Since the logarithmic function changes weakly with N_v / N_c, the Fermi energy of an intrinsic semiconductor is approximately at mid-gap. The temperature dependence of the intrinsic Fermi energy is weak due to the (weak) logarithmic dependence of the Fermi energy on the temperature. Using Boltzmann statistics the Fermi energy allows us to determine the ***intrinsic carrier concentration***, n_i, of electrons and holes in an undoped semiconductor.

$$\boxed{n_i = \sqrt{N_v \, N_c} \, \exp\left(-\frac{E_g}{2\,kT}\right)} \tag{14.10}$$

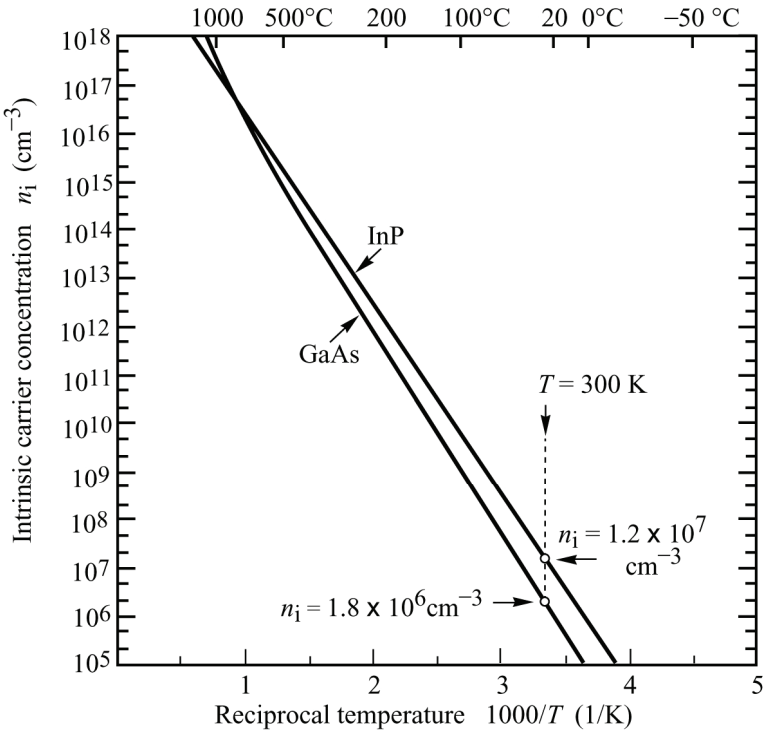


Fig. 14.2. Intrinsic carrier concentrations of GaAs and InP as a function of temperature. The slopes of the curves are proportional to the band-gap energy (after Thurmond, 1975; Laufer *et al.*, 1980).

According to this equation the intrinsic carrier concentration increases exponentially with temperature. In addition, the effective density of states have the comparatively weak temperature dependence of $N_{c,v} \propto T^{3/2}$. The intrinsic carrier concentration is of special importance. Calculating the product of electron and hole concentration for *any* (non-degenerate) Fermi level using Boltzmann statistics yields

$$n\ p \quad = \quad n_i^2 \quad = \quad N_v\ N_c\ \exp\left(-\frac{E_g}{kT}\right) . \tag{14.11}$$

Thus the product *n p* is a constant at a given temperature and, since the result does not depend on the Fermi level, is independent of the doping concentration. The intrinsic carrier concentrations of GaAs and InP are shown as a function of temperature in ***Fig***. 14.2.

14.2 Extrinsic semiconductors (single donor species)

Substitutional donors and acceptors have an excess or a deficit electron in their outer electron shell, respectively, as compared to the replaced lattice atom. ***Donors*** have one excess electron which can be *donated* to the conduction band. ***Acceptors*** have one less electron than the replaced lattice atom and can *accept* an electron from the filled valence band of the semiconductor, thereby creating a *hole*. Here we consider donors and acceptors being represented by an energy state close to the conduction band edge (donor) or close to the valence band edge (acceptor), as shown in ***Fig***. 14.3. We will next investigate the free carrier concentration as a function of temperature in a semiconductor with donor impurities of *one* chemical species. The donor concentration is assumed to be N_D.

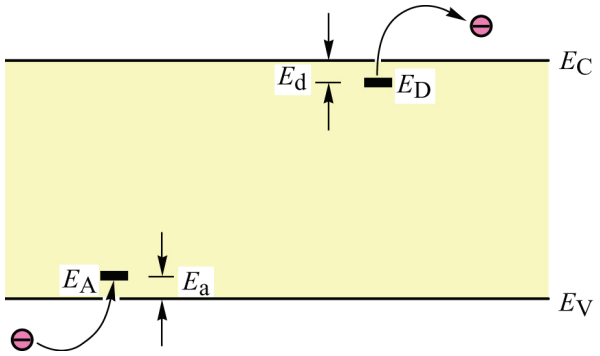


Fig. 14.3. Energy levels of acceptors and donors in the semiconductor band diagram. A donor (acceptor) level at energy E_D (E_A) has an ionization energy E_d (E_a).

The charge state of donors is *neutral* when occupied by an electron and positively charged if the electron is excited to the conduction band. The total concentration of donors is the sum of neutral donor concentration and ionized donor concentration, *i. e.*

$$N_D \quad = \quad N_D^0 \ + \ N_D^+ . \tag{14.12}$$

The energy of the donor impurity state is denoted as E_D. The donor energy is frequently given with respect to the conduction band edge, that is

$$E_d \quad = \quad E_C \ - \ E_D . \tag{14.13}$$

The probability of occupation of an acceptor or donor follows Fermi – Dirac statistics.

Consequently, the concentration of neutral donors, *i. e.* donors occupied by an electron is

$$N_D^0 = N_D \, f_F(E_D) \tag{14.14}$$

where $f_F(E_D)$ is the value of the Fermi–Dirac distribution at the energy of the donor. With $N_D^+ = N_D - N_D^0$ one obtains the concentration of ionized donors

$$\begin{aligned} N_D^+ &= N_D \left[1 - f_F(E_D)\right] = N_D \left[1 - \left(1 + \frac{1}{g} \exp\left(\frac{E_D - E_F}{kT}\right)\right)^{-1}\right] \\ &= N_D \left[1 + g \, \exp\left(\frac{E_F - E_D}{kT}\right)\right]^{-1} \end{aligned} \tag{14.15}$$

where g is the ground-state degeneracy of the donor. The value of the ground-state degeneracy in GaAs is $g = 2$ for hydrogen-like donors since the donor can donate one electron of either spin (see Chap. 1). The ground-state degeneracy of acceptors in GaAs is $g = 4$, since the acceptor can accept electrons of either spin from the heavy-hole and the light-hole valence band (see Chap. 1). Note that Eq. (14.15) is limited to concentrations below the Mott transition (see Chap. 1). Above the Mott transition, impurities cannot bind charge carriers, *i. e.* donors and acceptors cannot be in the neutral charge state.

If a semiconductor has one carrier type dominating due to doping, the other carrier type has an extremely small equilibrium concentration. If, for example, GaAs is doped with $N_D = 10^{17}\,\mathrm{cm}^{-3}$ donors and $n \approx 10^{17}\,\mathrm{cm}^{-3}$, the hole-concentration inferred from Eq. (14.10) at 300 K is $p = n_i^2 / n = 3.2 \times 10^{-3}\,\mathrm{cm}^{-3}$. Thus, there are approximately 3 holes in 1000 cm^3 of this n-type semiconductor. The very small concentration of the ***minority carrier*** allows us to completely neglect minority carriers in many semiconductor structures. Such semiconductor devices are called ***majority carrier devices***.

Charge neutrality is maintained in a doped semiconductor and has to be taken into account in addition to Fermi – Dirac statistics. Since minority carriers can be neglected, the free carrier concentration coincides with the ionized dopant concentration. If we restrict ourselves to n-type semiconductors, then

$$n = N_D^+ \, . \tag{14.16}$$

We now consider the semiconductor at low temperatures, when most electrons occupy donor states. Then Boltzmann statistics can be used for the occupation of conduction band states according to

$$n = N_c \exp\left(-\frac{E_C - E_F}{kT}\right) . \tag{14.17}$$

If Fermi–Dirac statistics are used for the occupation of the donor level according to Eq. (14.15), one obtains a quadratic equation for the free carrier concentration

$$n^2 - \frac{1}{g} N_D \, N_c \, e^{-E_d / kT} + \frac{1}{g} n \, N_c \, e^{-E_d / kT} = 0 \, . \tag{14.18}$$

At low temperatures the free carrier concentration, n, is much smaller than the donor

concentration, N_D. Thus, the third term of the quadratic equation is much smaller than the second term. The free carrier concentration is given by

$$n \approx \left(\frac{1}{g} N_D N_c\right)^{1/2} \exp\left(-\frac{E_d}{2kT}\right) \tag{14.19}$$

where the ground state degeneracy for donors is $g = 2$.

Equation (14.19) was first obtained by de Boer and van Geel (1935) by the method described here. The formulas can also be obtained by minimizing the free energy change due to thermal excitation of electrons from donor states to conduction band states (Mott and Gurney, 1940). At higher temperatures all donors become ionized. The carrier concentration is then constant $n = N_D^+ = N_D$ and independent of temperature. This temperature regime is called the ***saturation regime***.

As the temperature is increased even further, the *intrinsic* carrier concentration n_i increases and at sufficiently high temperatures assumes values comparable or higher than the dopant concentration. For most technologically useful semiconductors, the crossover from the saturation to the ***intrinsic regime*** occurs at temperatures much higher than room temperature. The three temperature regimes *(i)* thermal ionization regime *(ii)* saturation regime and *(iii)* intrinsic regime are shown schematically in ***Fig***. 14.4 along with the associated activation energies.

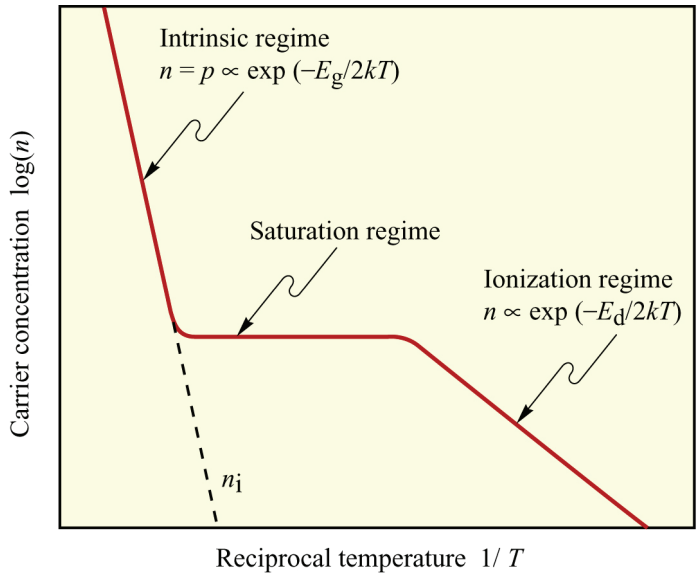


Fig. 14.4. Carrier concentration as a function of reciprocal temperature for an uncompensated n-type semiconductor. The donor is assumed to form one level at energy E_d below the conduction band edge. Three regimes, namely (*i*) the ionization regime, (*ii*) the saturation regime, and (*iii*) the intrinsic regime and their corresponding activation energies are indicated (after Smith, 1986).

The thermal ionization energy of a donor can be obtained from the slope of n versus reciprocal temperature. Rearrangement of Eq. (14.19) yields

$$E_d = -2k \frac{d(\ln n)}{d(1/T)} \tag{14.20}$$

which allows one to determine E_d directly from the temperature dependent carrier concentration. The change in carrier concentration with increasing temperatures also implies a continuously changing Fermi level in the semiconductor. Consider an n-type semiconductor. At low temperatures the donor levels are filled, while the conduction band is empty. Thus, the Fermi level must be slightly above the donor level. As the temperature increases, the Fermi distribution

becomes smeared out, as conduction band states become filled and donor states become unoccupied. Simultaneously the Fermi level moves deeper into the forbidden gap. At still higher temperatures, the Fermi level approaches the (near) mid-gap level and the semiconductor becomes intrinsic.

14.3 Extrinsic semiconductors (two donor species)

In the following, the free carrier concentration as a function of temperature is investigated in a semiconductor with two different species of donor impurities. It is assumed that the two donors form two different energy levels in the gap of the semiconductor. The two types of donor levels can originate from two different chemical species (*e. g.* Sn and Te donors in GaAs).

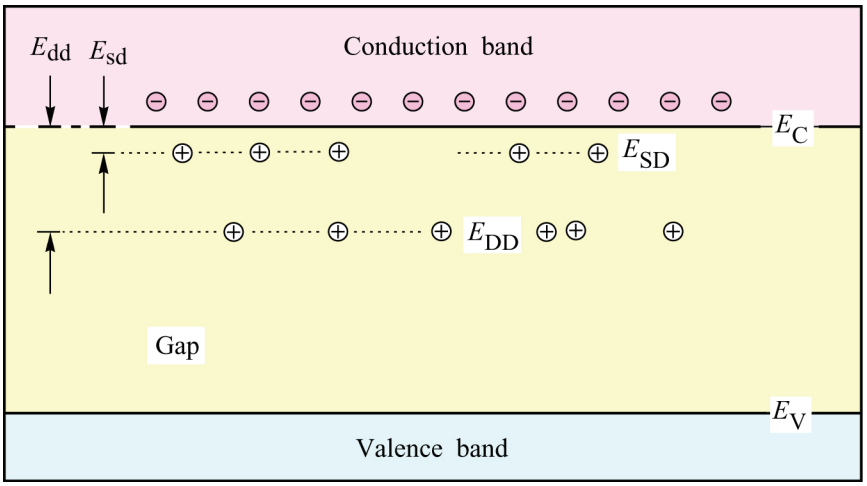


Fig. 14.5. Band diagram of a semiconductor with two species of donors namely a shallow donor with energy E_{SD} (ionization energy E_{sd}) and a deep donor with energy E_{DD} (ionization energy E_{dd}).

If both donor types have very similar thermal ionization energies, there is no necessity to differentiate between the two types of donors. It is therefore assumed that the two types of donors have markedly different ionization energies. In particular, we assume that one of the donors is relatively shallow and the other one is relatively deep. The band diagram of a semiconductor with two types of donors with energies E_{SD} and E_{DD} is shown in ***Fig***. 14.5. If the shallow and deep donor concentrations are given by N_{SD} and N_{DD}, respectively, then the free carrier concentration is given by

$$n = N_{SD}^{+} + N_{DD}^{+} . \tag{14.21}$$

The carrier concentration in the conduction band is given by the Boltzmann distribution

$$n = N_c \exp\left(-\frac{E_C - E_F}{kT}\right) \tag{14.22}$$

while Fermi–Dirac statistics is assumed for the donor levels (see Eq. 14.15).

$$N_{SD}^{+} = N_{SD} - \frac{N_{SD}}{1 + \frac{1}{g}\exp\left(\frac{E_{SD} - E_F}{kT}\right)} , \tag{14.23}$$

$$N_{DD}^{+} = N_{DD} - \frac{N_{DD}}{1 + \frac{1}{g}\exp\left(\frac{E_{DD} - E_F}{kT}\right)} \tag{14.24}$$

where E_{SD} and E_{DD} are the energies of the donor states. We assume that $E_{sd} << E_{dd}$ and $E_{dd} << E_g$, where $E_{sd} = E_C - E_{SD}$ and $E_{dd} = E_C - E_{DD}$.

All donors are neutral at very low temperatures. As temperature increases, shallow donors will donate their electrons to the conduction band, until all shallow donors are ionized. As the temperature is further increased, the deep donors start to become ionized until all deep donors are ionized. At even higher temperatures the intrinsic carrier concentration exceeds the dopant concentration and the semiconductor becomes intrinsic. In the following, the ionization regimes of the shallow and the deep donor are investigated.

At low temperatures when both types of donors are neutral, the Fermi energy is higher than the shallow donor energy. Then the energy difference between the Fermi energy and the deep donor energy is relatively large and according to Eq. (14.24), the deep donor can be considered as neutral, *i. e.* $N_{DD}^{+} = 0$. The carrier concentration is then given by

$$n = N_{SD}^{+} = N_{SD} - \frac{N_{SD}}{1 + \frac{1}{g}\exp\left(\frac{E_{SD} - E_F}{kT}\right)} \,. \tag{14.25}$$

Using Boltzmann statistics for the conduction band one obtains the quadratic equation

$$n^2 - \frac{1}{g} N_{SD}\, N_c \exp\left(-E_{sd}/kT\right) + \frac{1}{g} n\, N_c \exp\left(-E_{sd}/kT\right) = 0 \tag{14.26}$$

which is identical to the single donor equation Eq. (14.18). Thus, the low temperature solution is

$$\boxed{n \approx \left(\frac{1}{2} N_{SD}\, N_c\right)^{1/2} \exp\left(-\frac{E_{sd}}{2\,kT}\right)} \tag{14.27}$$

where the ground-state degeneracy is assumed to be $g = 2$. Ionization of the shallow donor continues until all shallow donors are ionized, *i. e.* $n = N_{SD}^{+} = N_{SD}$.

As the temperature is increased further, deep donors become ionized. Using Boltzmann statistics for the conduction band (Eq. 14.22) and Fermi–Dirac statistics for the deep donor (Eq. 14.24) one obtains the quadratic equation

$$(N_{DD}^{+})^2 + N_{DD}^{+}\left(N_{SD}^{+} + \frac{1}{g} N_c\, e^{-E_{dd}/kT}\right) - \frac{1}{g} N_{DD}\, N_c\, e^{-E_{dd}/kT} = 0\,. \tag{14.28}$$

For $N_{SD}^{+} >> (g^{-1})\, N_c \exp(-E_{dd}/kT)$ one obtains

$$(N_{DD}^{+})^2 + N_{SD}^{+}\, N_{DD}^{+} \approx \frac{1}{g} N_{DD}\, N_c\, e^{-E_{dd}/kT} \,. \tag{14.29}$$

Since the free carrier concentration is the sum of ionized deep and shallow donor concentration ($n = N_{SD}^{+} + N_{DD}^{+}$) the equation can be written as

$$n\,(n - N_{SD}^{+}) \;\approx\; \frac{1}{2}\,N_{DD}\,N_c\,e^{-E_{dd}\,/\,kT} \tag{14.30}$$

where the donor ground-state degeneracy is assumed to be $g = 2$.

At the elevated temperatures considered here, the shallow donor is ionized ($N_{SD}^{+} = N_{SD}$) and therefore the slope of the carrier density with respect to temperature follows the proportionality

$$\boxed{n\,(n - N_{SD}) \;\propto\; e^{-E_{dd}\,/\,kT}} \tag{14.31}$$

Note that this equation is significantly different from the simple relation $n \propto \exp(-E_d / kT)$, which would lead to incorrect results if applied to a semiconductor with two types of donors. Equation (14.31) was applied to shallow and deep Si donors in $Al_xGa_{1-x}As$ (Schubert and Ploog, 1984).

The ionization of the deep donor continues until shallow and deep donors are ionized, which corresponds to the saturation regime. At even higher temperatures the intrinsic carrier concentration increases above the dopant concentration and the semiconductor becomes intrinsic. The carrier concentration is shown in ***Fig***. 14.6 for a semiconductor containing two different donor species as a function of temperature. The different saturation and ionization regimes along with their activation energies are indicated in the figure. Note that the different ionization regimes discussed above may not be as clearly distinguishable if the difference between E_{sd} and E_{dd} is small.

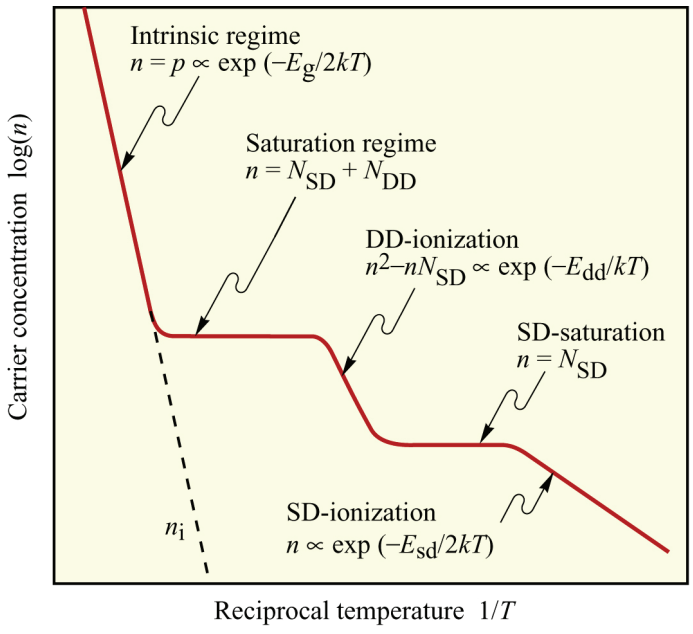


Fig. 14.6. Carrier concentration as a function of reciprocal temperature for a semiconductor with two different dopant species, namely a shallow donor (SD) and a deep donor (DD). At low temperatures the shallow donor becomes ionized and saturates when $n = N_{SD}$. At higher temperatures the deep donor becomes ionized and saturates ($n = N_{SD} + N_{DD}$). At even higher temperatures the semiconductor becomes intrinsic ($n = n_i$). The activation energies of each regime are indicated.

14.4 Compensated semiconductors

A partially compensated semiconductor contains dopant atoms of one type (n- or p-type) and, in addition, a smaller number of dopants of the other type. The band diagram of a partially compensated n-type semiconductor with a small number of acceptors is shown in ***Fig***. 14.7. The free carrier concentration is given by

$$n \;=\; N_D^{+} \;-\; N_A^{-}\;. \tag{14.32}$$

Electrons from donors prefer to occupy lower-energy acceptor states at all temperatures. Thus, even for low temperatures ($T \to 0\,\mathrm{K}$) *all* acceptors and *some* donors are ionized. Fermi–Dirac statistics for the occupation of the donor state (see Eq. 14.14) and Eq. (14.32) yield

$$n + N_A^- = N_D^+ = N_D - \frac{N_D}{1 + \frac{1}{g}\exp\left(\frac{E_D - E_F}{kT}\right)} . \qquad (14.33)$$

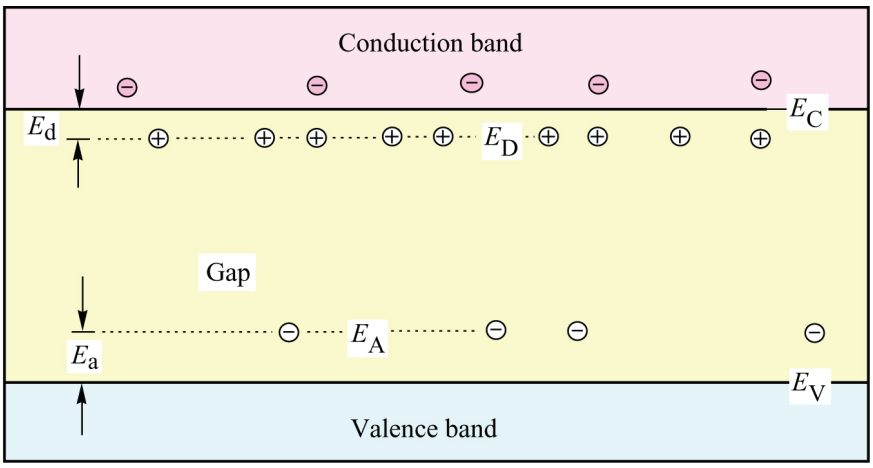


Fig. 14.7. Band diagram of a lightly compensated n-type semiconductor.

Using Boltzmann statistics for the free electron concentration in the conduction band allows one to eliminate the Fermi energy. One obtains the quadratic equation

$$n^2 + n\left(N_A + \frac{1}{g} N_c\, e^{-E_d/kT}\right) - \frac{1}{g} N_c\, e^{-E_d/kT}\,(N_D - N_A) = 0 \qquad (14.34)$$

where acceptors are assumed to be ionized at all temperatures, *i. e.* $N_A = N_A^-$. The solution of the quadratic equation is, for $g = 2$:

$$n = \frac{-1}{2}\left(N_A + \frac{N_c}{2} e^{-E_d/kT}\right) + \frac{1}{2}\left[\left(N_A + \frac{N_c}{2} e^{-E_d/kT}\right)^2 + 2N_c e^{-E_d/kT}(N_D - N_A)\right]^{\frac{1}{2}} \qquad (14.35)$$

At low temperatures when $(1/2)\,N_c \exp(E_d/kT) << N_A$ the equation simplifies to

$$n = -\frac{1}{2} N_A + \frac{1}{2} N_A \left[1 + 2\, N_c\, e^{-E_d/kT}\left(\frac{N_D - N_A}{{N_A}^2}\right)\right]^{1/2} . \qquad (14.36)$$

An approximate solution of the equation can be found by applying $(1 + x)^{1/2} \approx 1 + (1/2)\,x$ (for $x << 1$) to the square root term of the equation.

$$\boxed{n \approx \frac{1}{2} N_c \frac{N_D - N_A}{N_A} e^{-E_d/kT}} \qquad (14.37)$$

Note that the temperature dependence of the carrier concentration as a function of temperature has a different slope as compared with the uncompensated semiconductor (see Eq. 14.19). The slopes are different by a factor of two for the compensated and uncompensated case.

As the temperature is further increased, $N_D >> (1/2) N_c \exp(-E_d/kT) >> N_A$ and Eq. (14.35) simplifies to

$$n \approx \left(\frac{1}{2} N_c N_D\right)^{1/2} \exp\left(-\frac{E_d}{2kT}\right) \tag{14.38}$$

which is identical to the uncompensated case given by Eq. (14.19).

Even further increase of the temperature results in fully ionized donors. The free carrier concentration is then given by $n = N_D^+ - N_A^- = N_D - N_A$. The two different slopes for the carrier concentration vs. temperature have indeed been observed experimentally (Morin, 1959). The carrier concentration versus temperature of a compensated semiconductor is schematically illustrated in ***Fig***. 14.8, where the different regimes are indicated.

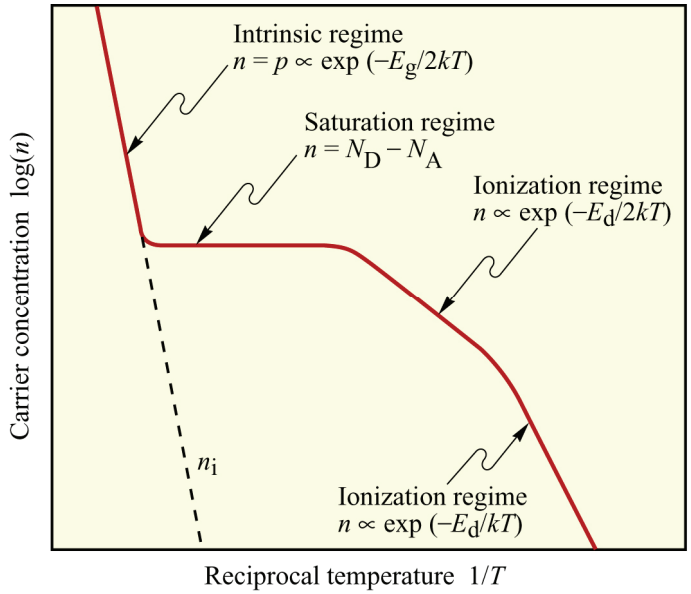


Fig. 14.8. Carrier concentration as a function of reciprocal temperature for a partially compensated n-type semiconductor. The ionization regime is characterized by two different activation energies, namely E_d and $E_d/2$.

Exercise 1: ***Approximate rules***. Answer the following questions.

(a) In a non-degenerately doped semiconductor, is the Fermi level is within the (i) valence band, (ii) forbidden gap, or (iii) conduction band?
(b) In a degenerately doped n-type semiconductor, is the Fermi level within the (i) valence band, (ii) forbidden gap, or (iii) conduction band?
(c) Where is the Fermi level located in an n-type semiconductor at $T \to 0$?
(d) Where is the Fermi level located in an n-type semiconductor for $T \to \infty$?
(e) What are the possible charge states of a donor?
(f) What are the possible charge states of an acceptor?
(g) States below the Fermi level have a probability greater 50% of being (i) occupied or (ii) unoccupied?
(h) States above the Fermi level have a probability greater 50% of being (i) occupied or (ii) unoccupied?
(i) What is the charge state of an occupied donor?

(j) What is the charge state of an unoccupied donor?
(k) What is the charge state of an occupied acceptor?
(l) What is the charge state of an unoccupied acceptor?
(m) The Fermi level in an intrinsic semiconductor is *approximately* but not *exactly* in the middle of the forbidden gap. Why?
(n) Semiconductor devices are generally operated in the (i) ionization, (ii) saturation, or (iii) intrinsic regime?
(o) Calculate the electron concentration in GaAs at room temperature using the Boltzmann and the Fermi–Dirac distribution for (i) $E_F = E_C - 2kT$, (ii) $E_F = E_C - kT$, and (iii) $E_F = E_C$. Determine the error obtained when using the Boltzmann distribution.

Solution:
(a) Forbidden gap
(b) Conduction band
(c) At the donor level
(d) Near middle of gap
(e) Neutral and positive
(f) Neutral and negative
(g) States below the Fermi level have a probability greater 50% of being occupied
(h) States above the Fermi level have a probability greater 50% of being unoccupied
(i) Neutral
(j) Positive
(k) Negative
(l) Neutral
(m) If the density of states in the conduction and valence band would be identical, then the Fermi level in the intrinsic case would be *exactly* in the middle of the forbidden gap. This is because the Fermi–Dirac distribution is symmetrical and the electron and hole concentrations are equal ($n = p = n_i$). However, because the density of states in the conduction and valence band are generally not equal, the Fermi level is slightly shifted from the mid-gap position towards the band with the lower density of states. This insures that $n = p = n_i$.
(n) Saturation regime
(o) In GaAs, $N_c = 4.4 \times 10^{17}$ cm^{-3}; At room temperature, $kT = 25$ meV

$E_F = E_C - 2kT$ ➔	$n_{Boltzmann} = 5.9 \times 10^{16}$ cm^{-3}	$n_{FD} = 5.7 \times 10^{16}$ cm^{-3}	Error = 3.8%
$E_F = E_C - kT$ ➔	$n_{Boltzmann} = 1.6 \times 10^{17}$ cm^{-3}	$n_{FD} = 1.4 \times 10^{17}$ cm^{-3}	Error = 12%
$E_F = E_C$ ➔	$n_{Boltzmann} = 4.4 \times 10^{17}$ cm^{-3}	$n_{FD} = 3.8 \times 10^{17}$ cm^{-3}	Error = 31%

References

de Boer J. H. and van Geel W. C. "Title unknown to EFS" *Phys.* **2**, 186 (1935)
Kittel C. and Kroemer H. *Thermal Physics*, 2nd edition (Freeman, San Francisco, 1980)
Lang D. V., Grimmeiss H. G., Meijer E., Jaros M. "Complex nature of gold-related deep levels in silicon" *Physical Review B***22**, 3917 (1980)
Laufer P. M., Pollack F. H. Nahory, R. E. and Pollack, M. A. "Electroreflectance investigation of $In_{1-x}Ga_xAs_yP_{1-y}$ lattice-matched to InP" *Solid State Communications* **36**, 419 (1980)
Morin F. F., in *Semiconductors*, edited by N. B. Hannay, p. 31 (Reinhold, New York, 1959)
Mott N. F. and Gurney R. W. *Electronic Processes in Ionic Crystals*, p. 156 (Oxford Univ. Press, Oxford, 1940)
Reif F., *Fundamental of Statistical and Thermal Physics* (McGraw-Hill, New York, 1965)
Schubert E. F. and Ploog K. "Shallow and deep donors in direct-gap n-type $Al_xGa_{1-x}As$: Si grown by molecular-beam epitaxy" *Physical Review B***30**, 7021 (1984)
Smith R. A., *Semiconductors* (Cambridge University Press, Cambridge UK, 1978)
Thurmond C. D. "Standard Thermodynamic functions for the formation of electrons and holes in Ge, Si, GaAs, and GaP" *Journal of the Electrochemical Society* **122**, 1133 (1975)
van Vechten J. A. and Thurmond C. D. "Entropy of ionization and temperature variation of ionization

levels of defects in semiconductors" *Physical Review B**14***, 3539 (1976a)
van Vechten J. A. and Thurmond C. D. "Comparison of theory with quenching experiments for the entropy and enthalpy of vacancy formation in Si and Ge" *Physical Review B**14,*** 3551 (1976b)

15

Impurities in semiconductors

15.1 Bohr's hydrogen atom model

Shallow impurities are of great technological importance in semiconductors since they determine the conductivity and the carrier type of the semiconductor. Shallow impurities are defined as impurities which are ionized at room temperature. This condition limits the ionization energy of such impurities to values << 100 meV. Shallow impurities can be either acceptors or donors, *i. e.* 'accept' electrons from the valence band or 'donate' electrons to the conduction band.

The hydrogen atom model can serve as the basis for the calculation of many properties of shallow impurities such as ionization energy and state wave functions. In this chapter, the hydrogen atom is analyzed in terms of Bohr's semi-classical model and in terms of a quantum mechanical approach. The hydrogen atom model is then applied to shallow impurities. Properties such as ionization energies, wave functions, central cell correction terms, and screening of impurity potentials by free carriers are summarized.

Impurities in semiconductors can be incorporated on substitutional sites, interstitial sites, or as impurity complexes. Here, we restrict ourselves to *substitutional, shallow* impurities. Examples for such impurities are Be, Zn, Si, and Sn. These impurities are shallow, *i. e.* their ionization energy is comparable to the thermal energy *kT* at room temperature. As a consequence, shallow impurities are fully ionized at room temperature. The hydrogen atom model has proven to predict accurately many properties of shallow impurities.

Coulomb potential

The electrostatic potential of a point charge is called the Coulomb potential or the 1 / *r* potential. The Coulomb potential of a positive point charge (+*e*) in vacuum located at $r = 0$ is obtained from Poisson's equation and is given in spherical coordinates by

$$V(r) = \frac{e}{4\pi\varepsilon_0 r} \tag{15.1a}$$

where *e* is the elementary charge and ε_0 is the permittivity of vacuum. Analogously, the Coulomb potential of a positively charged impurity located at $r = 0$ in a semiconductor with the dielectric constant $\varepsilon_r = \varepsilon / \varepsilon_0$ is given by

$$V(r) = \frac{e}{4\pi\varepsilon r} \tag{15.1b}$$

where ε is the permittivity of the semiconductor.

Binding energy and Bohr radius

The hydrogen atom model developed by Bohr is based on (*i*) classical mechanics of an electron

in the Coulomb potential of a positive point charge and on (*ii*) the quantization of the electron angular momentum. The Bohr model predicts many of the physical properties of the hydrogen atom most notably the emission spectra of the atom. The model is a fascinating example of the simplicity and the power of quantum mechanics. For the classical motion of an electron in the Coulomb potential, the potential energy is given by

$$E_{\text{pot}} = -\frac{e^2}{4\pi\varepsilon_0 r} . \tag{15.2}$$

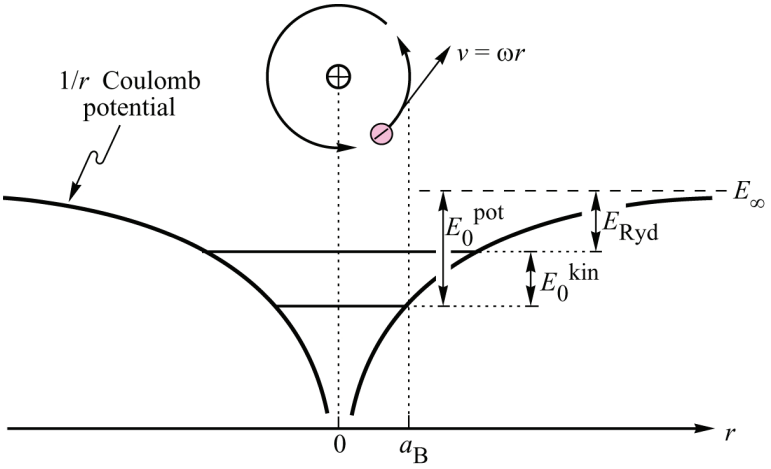


Fig. 15.1. Illustration of Bohr's hydrogen atom model. The electron orbits the positive proton with a $1/r$ Coulomb potential at the radius a_B. The ionization energy of the atom is the Rydberg energy $E_{Ryd} = E_\infty - |E_0^{pot}| + |E_0^{kin}|$.

For the hydrogen atom, the permittivity is that of vacuum since a vacuum is assumed between the proton and the electron. The schematic Coulomb potential and an electron orbiting the proton at a distance r are shown in ***Fig***. 15.1. The electrostatic Coulomb force F_C, acting on the proton and electron attracting them towards each other, is given by

$$F_C = e\mathcal{E}_C = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{r^2} \tag{15.3}$$

where $\mathcal{E}_C = dE_{pot}/dr$ is the Coulomb field. If the electron orbits the proton at a radius r and tangential velocity v, a centripetal force of magnitude F_z is required to keep the electron on the stationary orbit

$$F_z = \frac{m_0 v^2}{r} = m_0 r \omega^2 \tag{15.4}$$

where ω is the angular frequency of the electron. Equation (15.4) is valid, only if the electron mass, m_0, is much smaller than the proton mass, m_p. This condition is fulfilled, since $m_0 / m_p \approx 1 / 1840$. The first classical condition in Bohr's hydrogen atom model is $F_c = F_z$, *i. e.* the condition for a stationary circular motion of the electron around the proton.
The second condition for the Bohr atom is the quantization of the angular momentum of the electron which is given by

$$\boxed{m_0 v r = m_0 \omega r^2 = n\hbar} \qquad (n = 1, 2, 3 \ldots) \tag{15.5}$$

The validity of the angular momentum quantization can be visualized by recalling the wave character of the electron. An electron wave around the positive proton is shown in ***Fig***. 15.2. The electron wave is stable, only if the wave is interfering constructively with itself, *i. e.* when the length of the electron orbit equals integer multiples of the electron wavelength, λ,

$$2\pi r \quad = \quad n\lambda \qquad\qquad (n = 1, 2, 3 \ldots) \qquad\qquad (15.6)$$

The reader can easily verify that Eqs. (15.5) and (15.6) are identical by recalling that the kinetic energy of a particle is given by $E = \hbar^2 k^2 / 2m_0$ where $k = 2\pi / \lambda$ and m_0 are the electron wave vector and mass, respectively.

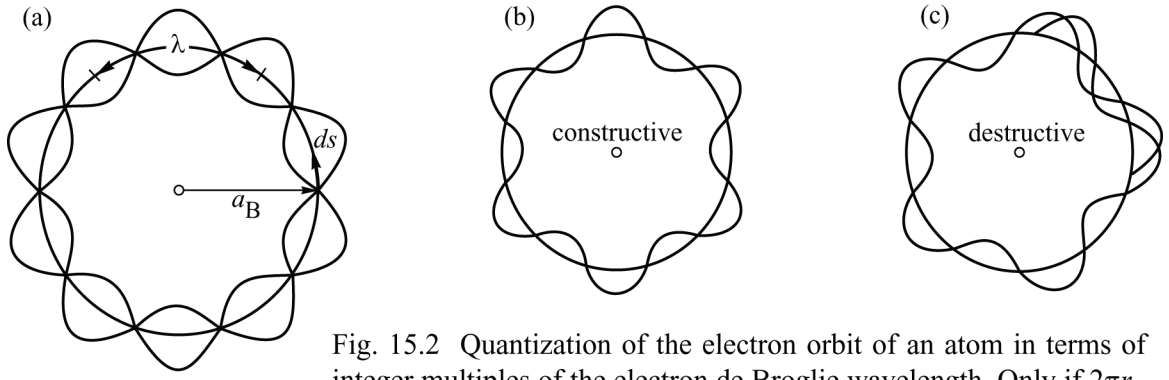


Fig. 15.2 Quantization of the electron orbit of an atom in terms of integer multiples of the electron de Broglie wavelength. Only if $2\pi r_B = n\lambda$, does constructive interference of the electron wave occur.

Elimination of force and velocity from Eqs. (15.3) – (15.5) yields the radii of the allowed electron orbits in the hydrogen atom

$$\boxed{a_{B,n} \quad = \quad \frac{4\pi\varepsilon_0 \, n^2 \, \hbar^2}{m_0 e^2}} \qquad\qquad (n = 1, 2, 3 \ldots) \qquad\qquad (15.7)$$

The radius of the ground state orbit ($n = 1$) is given by

$$a_B \quad = \quad 0.053 \text{ nm} \qquad\qquad (15.8)$$

which is called the ***Bohr radius*** of the hydrogen atom.

Insertion of the Bohr radii into Eq. (15.2) yields the potential energy of the electron

$$E_{pot,n} \quad = \quad \frac{-1}{(4\pi\varepsilon_0)^2} \frac{e^4 \, m_0}{n^2 \, \hbar^2} \qquad\qquad (n = 1, 2, 3 \ldots) \qquad\qquad (15.9)$$

Furthermore, the kinetic energy is obtained via the electron velocity of Eq. (15.5) according to

$$E_{kin,n} \quad = \quad \frac{1}{2} \frac{1}{(4\pi\varepsilon_0)^2} \frac{e^4 \, m_0}{n^2 \, \hbar^2} \qquad\qquad (n = 1, 2, 3 \ldots) \qquad\qquad (15.10)$$

Comparison of Eq. (15.9) with Eq. (15.10) reveals that the kinetic energy is just half of the

potential energy, *i. e.*

$$E_{\text{kin},n} = \frac{1}{2} \left| E_{\text{pot},n} \right| \tag{15.11}$$

The energy required to move the electron from the nth state energy, E_n, to the vacuum level at infinite distance from the proton, *i. e.* $E_\infty = E(r \to \infty)$, is given by

$$E_{\text{Ryd},n} = \frac{1}{2} \frac{1}{(4\pi\varepsilon_0)^2} \frac{e^4 m_0}{n^2 \hbar^2} \qquad (n = 1, 2, 3, ...) \tag{15.12a}$$

which, for $n = 1$, is called the ***Rydberg energy***. This energy is required to ionize a hydrogen atom. For $n = 1$ the Rydberg energy is given by

$$E_{\text{Ryd}} = 13.6 \text{ eV}. \tag{15.12b}$$

In the classical orbital motion of the electron around the proton, the ratio of the electron velocity and the velocity of light c can be calculated from the Bohr model. The ratio is obtained as

$$\alpha = \frac{v}{c} = \frac{e^2}{4\pi\varepsilon_0 \hbar c} \approx \frac{1}{137} \tag{15.13}$$

which is called ***the Sommerfeld fine structure constant***. Evaluation of Eq. (15.13) yields that the electron orbits the proton with a velocity of approximately 2200 km / s.

The magnetic field generated by the circular current of the orbiting electron can be calculated from Bohr's model using the Maxwell equations. It is given by

$$\mu_B = \frac{\mu_0 \hbar e}{2 m_0} \tag{15.14}$$

and is called the ***Bohr magneton***.

The above calculation demonstrates that the relatively simple Bohr model, *i. e.* classical mechanics and angular momentum quantization, provides many physical quantities of the hydrogen atom. The calculated state energies of the hydrogen atom were found to agree with hydrogen emission spectra. The Bohr model and its prediction of the electron energies was one of the first successes of the quantum theory. Further refinement of the model is obtained by considering not only circular orbits but also elliptical orbits. On such an elliptical orbit the velocity of the electron is a function of the position, *i. e.* the velocity is not constant as in the circular orbit. (The position dependence of the velocity is analogous to the planetary motion around the sun).

Using the momentum $p = \hbar k$ and $k = 2\pi / \lambda$, the angular momentum quantization condition, which for circular motion is given by Eq. (15.6), can be written for any orbit as

$$\frac{1}{2\pi} \int_0^{2\pi} p_\phi \, d\phi = n_\phi \hbar \qquad (n_\phi = 1, 2, 3 ...) \tag{15.15}$$

where $p_\phi = m \omega_\phi r^2$ is the position-dependent (that is angle-dependent) angular momentum. For a

circular orbit, the angular momentum is a constant and Eq. (15.15) reduces to Eq. (15.5).

The condition of classical mechanics for motion on an elliptical orbit and the angular momentum quantization condition lead to the total energy of the electron. The total energy is given by

$$E_{n_\phi} = \frac{1}{2}\frac{1}{(4\pi\varepsilon_0)^2}\frac{e^4 m_0}{n_\phi^2 \hbar^2} \qquad (n_\phi = 1, 2, 3 \ldots) \qquad (15.16)$$

which is identical to Eq. (15.12). Thus, elliptical orbits for the electron exist and have the same energy as electrons on circular orbits. The total energy of a particle on an elliptical orbit in a $1/r$ potential can be calculated by classical mechanics and depends only on the main axis a of the ellipsis. The main axis is then given by

$$a = \frac{4\pi\hbar^2 \varepsilon_0}{m_0 e^2} n_\phi^2 \, . \qquad (15.17)$$

The angular momentum of the particle on the elliptical orbit is given by

$$p_\phi^2 = \frac{a(1-\varepsilon^2) e^2 m_0}{4\pi\varepsilon_0} \qquad (15.18a)$$

where ε is the eccentricity of the ellipsis. Since the angular momentum is quantized according to $p_\phi = n_\phi \hbar$, one obtains with Eq. (15.18a)

$$n_\phi^2 \hbar^2 = \frac{a(1-\varepsilon^2) e^2 m_0}{4\pi\varepsilon_0} \, . \qquad (15.18b)$$

Inserting the main axis a, given by Eq. (15.17), into Eq. (15.18b) yields

$$\frac{b}{a} = \frac{n_\phi}{n} \qquad (15.19)$$

where a and b are the axes of the ellipsis. The angular momentum quantum number n_ϕ can assume values of 1, 2, 3 …, n, which represents a family of ellipses. If the angular quantum number coincides with the principal quantum number, *i. e.* $n_\phi = n$, the previously calculated circular orbit is obtained. The ellipses for the $n = 1$, 2, and 3 states are shown in ***Fig***. 15.3(a).

If the angular momentum quantum number is formally introduced as $l = n_\phi - 1$, then l can assume values of

$$l = 0, 1, 2 \ \ldots \ n-1 \, . \qquad (15.20)$$

The value of $l = 0$ corresponds to the ellipsis with the largest eccentricity. The value of $l = n - 1$ represents the circular orbit.

Each orbit of an electron around the proton of a hydrogen atom is fully determined by the principle quantum number, n, and the angular quantum number, l. The orbits of the electron for different quantum numbers and the corresponding energies are shown in ***Fig***. 15.3. Frequently,

the $l = 0$, 1, 2 and 3 orbitals are historically denoted as the s, p, d and f orbitals (White, 1934). If several quantum states have the same energy the states are called ***degenerate states***. For example, the two states determined by $n = 2$, $l = 0$ and $n = 2$, $l = 1$ are degenerate. In addition to the principal quantum number and the angular momentum number, the quantum number m describes the quantization of the azimuthal angular momentum in units of $\hbar$. For a discussion of the azimuthal quantum number, we refer to the literature (see, for example, Bohm, 1951).

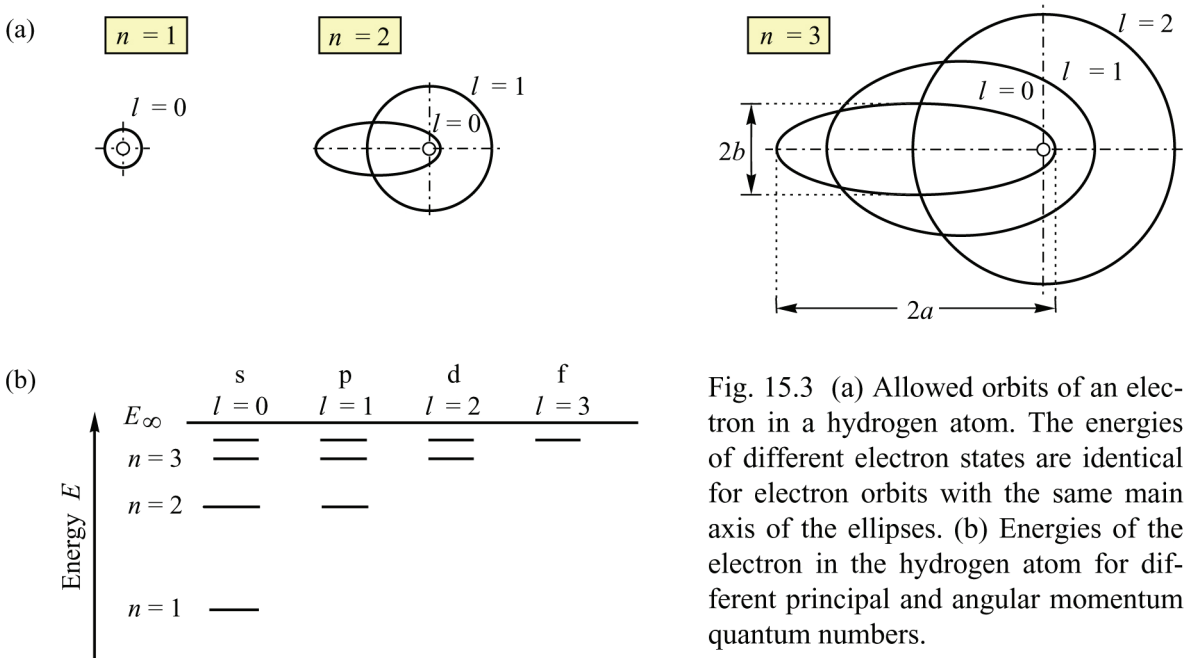


Fig. 15.3 (a) Allowed orbits of an electron in a hydrogen atom. The energies of different electron states are identical for electron orbits with the same main axis of the ellipses. (b) Energies of the electron in the hydrogen atom for different principal and angular momentum quantum numbers.

Upon any perturbation of the hydrogen atom, different electron orbits respond in different ways to the perturbation. Therefore, degenerate electron states will split and become non-degenerate upon a suitable perturbation. The perturbation of the hydrogen atom can be achieved, for example, by an electric field (Stark effect) or a magnetic field (Zeeman effect).

Even though the Bohr model explains many characteristics of the hydrogen atom it is limited in its applicability. For example, if the principles of the Bohr model are applied to the helium atom, incorrect results are obtained for the energy levels in that atom. In addition, the deterministic Bohr hydrogen model violates the quantum mechanical uncertainty principle. That is, momentum and position of the electron are exactly determined at all times in Bohr's model, which contradicts $\Delta x \, \Delta p \approx \hbar$. Nevertheless, due to its simplicity and clarity, the Bohr model has not lost its attractiveness.

The Bohr model was refined in 1925 by inclusion of the electron spin. The spin of an electron is also called its intrinsic angular momentum and can be visualized as the rotation of an electron around its own symmetry axis. Goudsmit and Uhlenbeck postulated the spin when conducting Zeeman effect experiments on hydrogen. Due to angular momentum quantization the difference between intrinsic angular momenta is $\hbar$. The intrinsic angular momentum is then given by

$$p_s = s\,\hbar \qquad \text{with} \qquad s = \pm 1/2 \tag{15.21}$$

where s is the spin quantum number which can assume values of $s = \pm 1/2$. If an electron with spin is subjected to an external magnetic field, the spin-axis will orientate (align) itself in a

parallel or antiparallel manner, as shown in ***Fig***. 15.4. All other orientations are of transient nature since the repulsive and attractive magnetic forces of the spin and external field tend to reorient the spin to the parallel or antiparallel orientation.

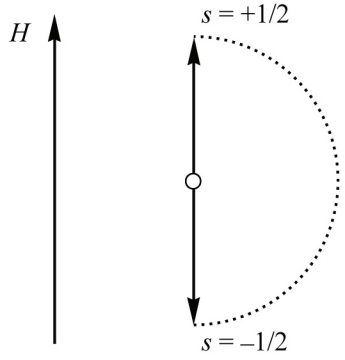


Fig. 15.4. Alignement of the spin of an electron in an external magnetic field H. The spin is either parallel (spin "up") or antiparallel (spin "down") to the magnetic field, as represented by the spin quantum number $s = \pm 1/2$. The difference in angular momentum between the two orientations is $h/(2\pi)$.

Wave functions of the hydrogen atom

The Bohr model predicts the energy levels of the hydrogen atom with amazing accuracy. However, the wave functions ψ_{nlm} and the probability distributions $\psi_{nlm}\,\psi_{nlm}^*$ cannot be obtained from the Bohr model. The Schrödinger equation must be solved in order to obtain the exact solutions. Although the exact solution of the hydrogen atom model goes beyond the scope of this book, the results of the exact hydrogen atom are summarized in the following. For further study, the reader is referred to textbooks on quantum mechanics (Bohm, 1951; Sherwin, 1959; Davydov, 1965; Borowitz, 1967; Saxon, 1968; Merzbacher, 1970).

The wave functions of the hydrogen atom can be obtained by solving the Schrödinger equation. For a particle with mass m and charge e in a potential V, the Schrödinger equation is given by

$$\frac{\hbar^2}{2\,m}\,\Delta\Psi \;+\; e\,V\,\Psi \;=\; -\frac{\hbar}{\mathrm{i}}\,\frac{\partial}{\partial t} \tag{15.22}$$

where $\Psi = \Psi(x,\ y,\ z,\ t) = \psi(x,\ y,\ z)\,\mathrm{e}^{\mathrm{i}\omega t}$ is the time-dependent wave function and $\Delta = (\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2)$ is the delta operator. Not being interested in the time dependence of the solution, we use the kinetic energy operator $E = -(\hbar/\mathrm{i})(\partial/\partial t)$ to obtain the time-independent Schrödinger equation

$$\boxed{-\frac{\hbar^2}{2\,m}\,\Delta\psi \;+\; (eV - E)\,\psi \;=0} \tag{15.23}$$

where $\psi = \psi(x,\ y, z)$ is the time-independent wave function. The Coulomb potential is a spherically symmetric potential and can be expressed solely as a function of the radius r, *i. e.* $V = V(r)$ (see Eq. (15.1)). It is useful to convert the Δ-operator into spherical coordinates (r, θ, ϕ), *i. e.*

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$
$$= \frac{1}{r^2}\,\frac{\partial}{\partial r}\left(r^2\,\frac{\partial}{\partial r}\right) + \frac{1}{r^2\,\sin\theta}\,\frac{\partial}{\partial \theta}\left(\sin\theta\,\frac{\partial}{\partial \theta}\right) + \frac{1}{r^2\,\sin^2\theta}\,\frac{\partial^2}{\partial \phi^2} \qquad (15.24)$$

where r, θ and ϕ are the radius, the polar angle, and azimuthal angle, respectively. The Schrödinger equation is a separable linear differential equation and can be solved by employing the product method. The wave functions can then be written as

$$\psi(r,\theta,\phi) = R(r)\;\Theta(\theta)\;\Phi(\phi)\;. \qquad (15.25)$$

Since $\psi\,\psi^*$ is the quantum mechanical probability density of the particle, the wave function ψ must satisfy the condition

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi\,\psi^*\,dx\,dy\,dz = \int_0^{\infty}\int_0^{2\pi}\int_0^{\pi} \psi\,\psi^*\,r^2\,\sin\theta\,d\theta\,d\phi\,dr = 1\;. \qquad (15.26)$$

The solution of the Schrödinger equation in spherical coordinates for the Coulomb potential is a set of orthogonal functions which are usually classified by the four quantum numbers n, l, m_l, and s. The quantum numbers are

n = the principal quantum number
l = the orbital angular momentum number
m_l = the azimuthal quantum number
s = the spin quantum number

The quantum number can assume values of

$$n = 1, 2, 3\;... \qquad (15.27a)$$

$$l = 0, 1, 2\;\;...\;\;n-2, n-1 \qquad (15.27b)$$

$$m_l = -l, -l+1\;\;...\;\;l-1, +l \qquad (15.27c)$$

$$s = -\tfrac{1}{2},\; +\tfrac{1}{2}\;. \qquad (15.27d)$$

Two electrons with different spin can occupy an orbit defined by the three quantum numbers n, l, and m_l (Pauli principle).

The wave functions corresponding to the three quantum numbers n, l, and m_l are designated as ψ_{nlm_l}. Correspondingly, the radial parts of the wave functions are denoted as R_{nlm_l} (see Eq. 15.25). The wave functions for some of the lowest states of the hydrogen atom (Sherwin, 1959) are given by

$$\psi_{100} = \frac{1}{\sqrt{\pi}}\left(\frac{1}{a_B}\right)^{3/2} e^{-r/a_B} \qquad (15.28a)$$

$$\psi_{200} = \frac{1}{4\sqrt{2\pi}}\left(\frac{1}{a_B}\right)^{3/2}\left(2 - \frac{r}{a_B}\right)e^{-r/2a_B} \tag{15.28b}$$

$$\psi_{210} = \frac{1}{4\sqrt{2\pi}}\left(\frac{1}{a_B}\right)^{3/2}(\cos\theta)\frac{r}{a_B}e^{-r/2a_B} \tag{15.28c}$$

$$\psi_{211} = \frac{1}{4\sqrt{2\pi}}\left(\frac{1}{a_B}\right)^{3/2}\frac{e^{i\phi}}{\sqrt{2}}(\sin\theta)\frac{r}{a_B}e^{-r/2a_B} \tag{15.28d}$$

$$\psi_{21-1} = \frac{1}{4\sqrt{2\pi}}\left(\frac{1}{a_B}\right)^{3/2}\frac{e^{-i\phi}}{\sqrt{2}}(\sin\theta)\frac{r}{a_B}e^{-r/2a_B}\ . \tag{15.28e}$$

Further solutions of the hydrogen atom state can be found in the literature (Bohm, 1951). The radial parts R_{nl0} ($m_l = 0$) for some of the lowest hydrogen atom states are shown in ***Fig***. 15.5. States of 's-type' (*i. e.* $l = 0$) symmetry have a maximum of the wave function at $r = 0$. States of 'p-type' (*i. e.* $l = 1$) symmetry have a node at $r = 0$.

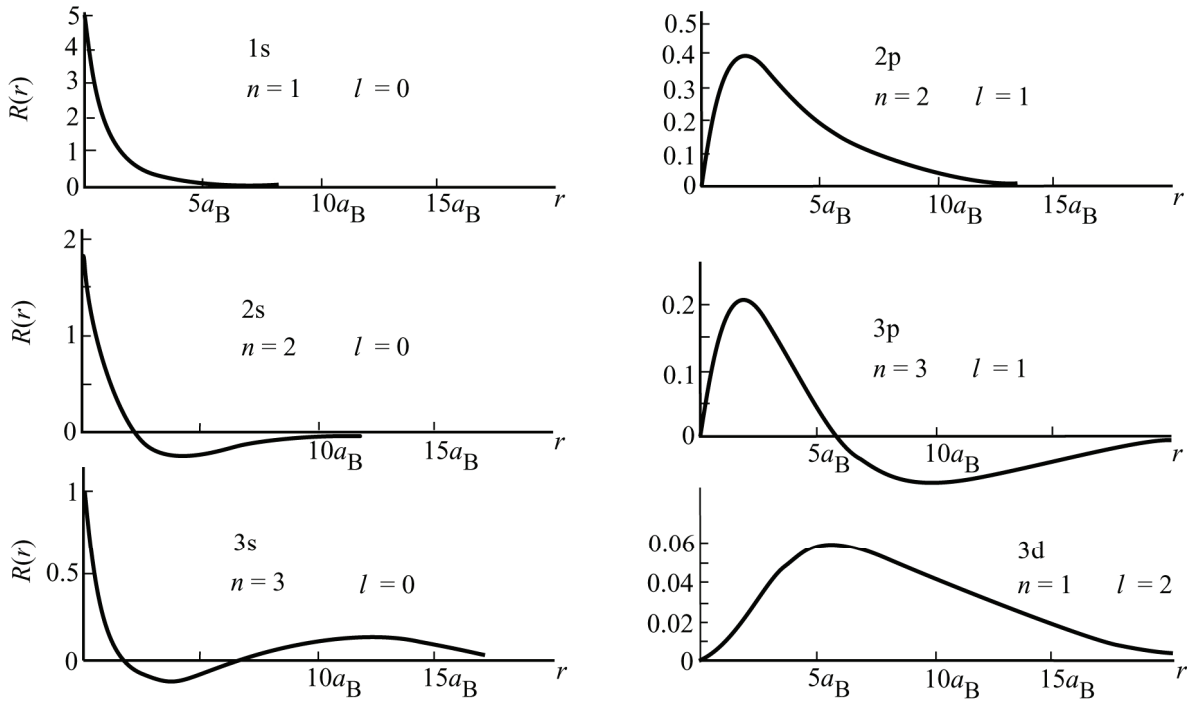


Fig. 15.5. Calculated radial parts of the wave functions for some elementary states of the hydrogen atom.

The probability of finding an electron at radius r can be obtained by integration over all angles θ and ϕ

$$p(r)\,\mathrm{d}r \;=\; \int_0^{2\pi}\int_0^{\pi} \psi\,\psi^*\,r^2\,\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi\,\mathrm{d}r\;. \tag{15.29}$$

For wave functions of s-type symmetry, ψ does not depend on θ and ϕ. Eq. (15.29) then yields

$$p(r) \;=\; \psi\,\psi^*\,4\pi\,r^2\;. \tag{15.30}$$

Using $\psi = \psi_{100}$ according to Eq. (15.28) one obtains

$$p(r) \;=\; 4\,\frac{r^2}{a_\mathrm{B}^3}\,\mathrm{e}^{-2r/a_\mathrm{B}}\;. \tag{15.31}$$

The probability $p(r)$ has a maximum at $r = a_\mathrm{B}$. Thus, the classical Bohr radius is the radius of maximum probability in the quantum mechanical picture.

15.2 Hydrogenic donors

The hydrogen atom model can be applied to shallow donors in III–V semiconductors. Properties predicted by the hydrogen atom model agree amazingly well with experimentally determined properties of shallow donors. Such donors are called ***effective-mass-like donors, hydrogen-like donors***, or briefly ***hydrogenic donors***. Ionization energy, wave functions, and effective Bohr radius are well predicted for such donors. The similarity of the hydrogen atom and donor impurities originates in the $1 / r$ coulombic potential of both entities.

Two modifications are required in order to apply Bohr's hydrogen atom model to shallow donors. These corrections are related to the effective mass of carriers and of the dielectric constant of the semiconductor (Bethe, 1942). First, the effective mass of electrons in semiconductors, m_e^*, differs from the electron mass quite significantly. The dispersion relation of a semiconductor with a spherically symmetric band structure is given by

$$E \;=\; \frac{\hbar^2\,k^2}{2\,m_\mathrm{e}^*}\;. \tag{15.32}$$

Thus, the dispersion relation $E(k)$ allows one to determine the effective mass according to

$$m_\mathrm{e}^* \;=\; \hbar^2\left(\frac{\mathrm{d}^2E}{\mathrm{d}k^2}\right)^{-1}\;. \tag{15.33}$$

The dispersion relation of Eq. (15.32) differs from the dispersion relation of a free electron just by the magnitude of the electron mass. In order to apply the hydrogen atom model, the electron mass must be replaced by the effective electron mass. The second correction arises from the dielectric properties of semiconductors. The Coulomb potential of a positive point-charge in a semiconductor located at $r = 0$ is given by

$$V(r) \;=\; \frac{e}{4\pi\,\varepsilon\,r} \tag{15.34}$$

which differs from Eq. (15.1a) by the static dielectric constant $\varepsilon_\mathrm{r} = \varepsilon / \varepsilon_0$. How does the potential

change in the presence of an electron orbiting the impurity charge? The polarization of the valence electrons is then more complicated and cannot be taken into account by the substitution of ε for ε_0. To answer the question we first make the simplifying assumption that the electronic charge can be described by a diffuse electron cloud with a spatial extent much larger than the lattice constant. In the limit of an infinitely large electron cloud, the potential of the positive impurity is correctly described by Eq. (15.34). The situation changes, however, if the electron were to orbit the impurity atom with a radius comparable to the lattice constant. In this case the polarization of the lattice depends on the donor as well as on the electron charge. The true potential is then not given by the dielectrically screened potential of Eq. (15.34). For such a small electron orbit, the polarization of lattice atoms is overestimated by Eq. (15.34). A smaller dielectric constant $\varepsilon_r^* < \varepsilon_r$ can be used to account for the reduction in polarization. It should therefore be noted that Eq. (15.34) assumes a Coulomb potential screened by the dielectric properties of the semiconductor (*i. e.* by polarization of tightly bound valence electrons and nuclei of the lattice) and that the equation can only be used if the electron can be described by a diffuse electron cloud with a large spatial extent. The true potential $V(r)$ which arises from the positive donor ion, the electron bound to the donor ion, and the polarization of the surrounding semiconductor is rather complicated and cannot be expressed in terms of a simple $1/r$ potential (Kohn, 1957a, 1957b). Employing the approximate $1/r$ potential, the Schrödinger equation is given by

$$-\frac{\hbar^2}{2\,m_e^*}\Delta\psi + \frac{e}{4\pi\varepsilon\, r}\psi = E\,\psi \tag{15.35}$$

which is called the effective-mass equation for a hydrogenic impurity.

Using the effective-mass and the dielectric constant corrections, the following properties of hydrogenic impurities can be derived. The ***effective Bohr radius*** is obtained from Eq. (15.7) and is given by

$$a_{B,n}^* = \frac{4\pi\varepsilon\, n^2\,\hbar^2}{m_e^* e^2} \qquad (n = 1, 2, 3\ \ldots) \tag{15.36}$$

The radius of the donor ground state ($n = 1$) is then given by

$$\boxed{a_B^* = \frac{4\pi\varepsilon\,\hbar^2}{m_e^*\, e^2} = \frac{\varepsilon_r}{m_e^*/m_0}\,a_B = \frac{\varepsilon_r}{m_e^*/m_0}\,0.53\ \text{Å}} \tag{15.37}$$

The effective Bohr radius is also called the *donor Bohr radius*. As an example, we consider a hydrogenic donor in GaAs with $\varepsilon_r = 13.1$ and $m_e^* = 0.067\ m_0$. Insertion of these values into Eq. (15.37) yields $a_B^* = 103$ Å which is the effective Bohr radius of donors in GaAs.

The ***effective Rydberg energy*** is obtained by applying the effective-mass and the dielectric constant corrections to Eq. (15.12).

$$E_{Ryd,n}^* = \frac{1}{2}\,\frac{1}{(4\pi\varepsilon)^2}\,\frac{e^4\, m_e^*}{n^2\,\hbar^2} \qquad (n = 1, 2, 3\ \ldots) \tag{15.38}$$

The ***donor ionization energy*** is required for a transition from $n = 1$ to $n \to \infty$ and is given by

$$E_\mathrm{d} = \frac{e^4 m_e^*}{2\,(4\pi\varepsilon\hbar)^2} = \frac{m_e^*/m_0}{\varepsilon_\mathrm{r}^2} E_\mathrm{Ryd} = \frac{m_e^*/m_0}{\varepsilon_\mathrm{r}^2} 13.6\,\mathrm{eV} \tag{15.39}$$

The donor ionization energy is occasionally also referred to as donor Rydberg energy. As an example, we consider a hydrogenic donor in GaAs and obtain $E_\mathrm{d} = 5.3$ meV which is in agreement with experimental results.

Finally, the wave functions of hydrogenic donors can be obtained from Eq. (15.28) by substituting the effective Bohr radius for the Bohr radius. The ground-state envelope wave function is then obtained as

$$\psi_{100}(r) = \frac{1}{\sqrt{\pi}} \left(\frac{1}{a_\mathrm{B}^*}\right)^{3/2} \mathrm{e}^{-r/a_\mathrm{B}^*} . \tag{15.40}$$

It should be noted that Eq. (15.40) describes the donor *envelope* function rather than the donor wave function. The actual donor ground-state wave function is given by (Kohn, 1957a)

$$\psi_{d,100}(\mathbf{r}) = \psi_{100}(r)\; u_k(\mathbf{r}) \tag{15.41}$$

where $u_k(\mathbf{r})$ is the lattice-periodic factor of the well-known Bloch function of conduction band electrons. The function $u_k(\mathbf{r})$ has translational symmetry with respect to the semiconductor lattice constant. The ground-state wave function according to Eq. (15.41) is schematically shown in ***Fig***. 15.6. The dashed curve represents the impurity envelope function of Eq. (15.40).

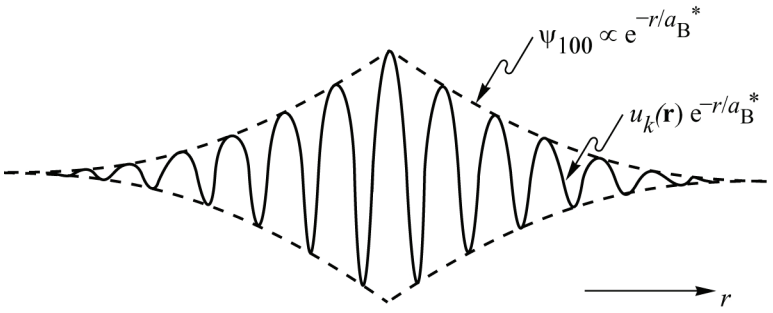


Fig. 15.6. Schematic illustration of a hydrogenic donor wave function with the quantum numbers $n = 1$ and $l = m_l = 0$. The wave function is the product of the lattice-periodic Bloch function $u_k(\mathbf{r})$ and the envelope function ψ_{100}.

It should be noted that the use of Bohr's hydrogen atom model for shallow impurities is not self-sufficient. The *ab initio* assumption of a 'large' electron cloud and the substitution of effective electron mass and dielectric constant cannot be justified solely on the basis of the hydrogenic model. Even though the hydrogenic model yields a relatively large electron orbit, this result does not justify the initial assumptions. However, more rigorous calculations (Kohn 1957a, 1957b; Madelung, 1978; Altarelli and Bassani, 1982) indeed demonstrate that the electron distribution has a spatial extent much larger than the lattice constant. The substitution of the electron mass, m_0, by the effective mass, m_e^*, is therefore justified since the electron orbit around the donor extends over many lattice constants. The effective mass of electrons in semiconductors is a direct consequence of the periodic potential of the lattice. Thus, electrons bound to donors are subject to the periodic potential, since the effective Bohr radius is much larger than the lattice constant, $a_\mathrm{B}^* >> a_\mathrm{B}$. If, in contrast, electrons were tightly bound ($a_\mathrm{B}^* \approx a_\mathrm{B}$) the effective-mass correction could not be applied. Similar arguments apply to the substitution of

the permittivity of the semiconductor, ε, for the permittivity of vacuum, ε_0. The lattice atoms are polarized by the Coulomb field which results in its reduction as compared to the field without polarization. The effect of the polarization is taken into account via the dielectric constant. Since the effective Bohr radius extends over many lattice constants, the use of the dielectrically screened Coulomb potential is justified. Despite the simplicity of the hydrogen atom model for shallow donors, the model yields quite accurate results.

The ***degeneracy*** of the donor ground state is a quantity required for the occupancy probability of the donor state (see Chap. 3). The ground state has the quantum numbers $n = 0$, $l = 0$, $m_l = 0$ and $s = \pm 1/2$. Thus, since the donor ground state can be occupied by an electron with spin + 1/2 or – 1/2, the ground state degeneracy is $g = 2$.

Exercise 1: ***Dopant ionization energies***. In GaAs, the dielectric constant is $\varepsilon_r = 13.1$ and the effective electron mass is $m_e^* = 0.067\ m_0$. In Si, the dielectric constant is $\varepsilon_r = 11.9$ and the effective electron mass is $m_e^* = 0.98\ m_0$. The experimental values for donor ionization energies for Si donors in GaAs and As donors in Si are 6 meV and 54 meV, respectively. Calculate the donor ionization energies and effective Bohr radii for donors in the two materials material by using the hydrogen model. Are hydrogenic theory and experimental values for the ionization energy in reasonable agreement?

Compare the donor ionization energies in GaAs and Si with the thermal energy *kT* at room temperature and at 77 K. Would you expect donors in the two materials to be ionized at 77 K? Would you expect donors in the two materials to be ionized at 300 K?

The hydrogen atom model can be also applied to *acceptors*, by using the *effective hole mass* rather then the effective electron mass. In GaAs, the dielectric constant is $\varepsilon_r = 13.1$ and the effective heavy-hole mass is $m_{hh}^* = 0.45\ m_0$. In GaN, the dielectric constant is $\varepsilon_r = 9.0$ and the effective heavy-hole mass is $m_{hh}^* = 0.8\ m_0$. The experimental values for acceptor ionization energies for Be acceptors in GaAs and Mg acceptors in GaN are 26 meV and 200 meV, respectively. Calculate the ionization energies and the effective Bohr radii for acceptors in the two materials by using the hydrogen model. Are hydrogenic theory and experimental values for the ionization energy in reasonable agreement?

Compare the ionization energies for acceptors in GaAs and acceptors in GaN with the thermal energy *kT* at room temperature. Are hydrogenic acceptors in GaAs mostly *ionized* or *neutral* at room temperature? Are hydrogenic acceptors in GaN mostly *ionized* or *neutral* at room temperature?

What are "shallow" dopants and why is it important that dopants are shallow?

Solution:

Given: GaAs: $\varepsilon_r = 13.1$; $m_e^* = 0.067\ m_0$; $E_d = 6$ meV (Si donor)
Si: $\varepsilon_r = 11.9$; $m_e^* = 0.98\ m_0$; $E_d = 54$ meV (As donor)

Using Eq. (15.39), $E_d = [(m_e^* / m_0) / \varepsilon_r^2]\ 13.6$ eV and Eq. (15.37), $\alpha_B^* = [\varepsilon_r / (m_e^*/m_0)]\ 0.53$ Å, we obtain:

For Si in GaAs:
Donor ionization energy $E_d = 5.3$ meV Effective Bohr radius $a_B^* = 103.6$ Å
For As in Si:
Donor ionization energy $E_d = 94$ meV Effective Bohr radius $a_B^* = 6.4$ Å

This result shows that for Si in GaAs, the hydrogen model is a good approximation. However the model does not work well for As donors in Si.

The larger mass in Si causes effective Bohr radius to be smaller, which means that the electron is closer to the donor atom core, so that a_B^* is not much larger than lattice constant. Thus hydrogen model does not work well.

Let us compare the donor ionization energy in GaAs and Si with the thermal energy kT at 77 K and 300 K.

At $T = 77$ K, the thermal energy is given by $kT = 6.63$ meV
At $T = 300$ K, the thermal energy is given by $kT = 25.9$ meV

Comparing the thermal energy with the ionization energies in GaAs and Si, it can be concluded that Si in GaAs will be ionized at 77 K, but that As in Si will not be ionized. At 300 K, Si donors in GaAs will be fully ionized, whereas As donors in Si may not be fully ionized. (This conclusion is solely based on $E_d > kT$. However, due to the high density of states in the Si conduction band, it turns out that most of As donors are in fact ionized.)

Given:					
	GaAs	$\varepsilon_r = 13.1$;	$m_{hh}^* = 0.45\ m_0$;	$E_a = 26$ meV	(Be acceptor)
	GaN	$\varepsilon_r = 9.0$;	$m_{hh}^* = 0.8\ m_0$;	$E_a = 200$ meV	(Mg acceptor)

For acceptor in GaAs

$E_a = [(m_{hh}^*/m_0) / \varepsilon_r^2]\ 13.6$ eV $= 35.6$ meV $\qquad a_B^* = [\varepsilon_r / (m_{hh}^* / m_0)]\ 0.53$ Å $= 15.4$ Å

For acceptor in GaN

$E_a = [(m_{hh}^*/m_0) / \varepsilon_r^2]\ 13.6$ eV $= 134$ meV $\qquad a_B^* = [\varepsilon_r / (m_{hh}^* / m_0)]\ 0.53$ Å $= 5.96$ Å

We see that the theoretical results are close to the experimental values.

At $T = 300$ K, $kT = 25.9$ meV. Thus at 300 K, acceptors in GaAs will be definitely ionized, but acceptors in GaN will not.

"Shallow" dopants are dopants that have a low ionization energy (on the order of kT at 300 K or lower), so that at room temperature all dopants are ionized. This is important, because full dopant activation at room temperature is desired.

15.3 Hydrogenic acceptors

The application of the hydrogenic model to acceptors in III–V semiconductors is complicated by their degenerate valence band structure. For hydrogenic donors, the effective electron mass was substituted for the electron mass. The substitution was possible, since the conduction band was assumed to be parabolic, isotropic, and non-degenerate (as for most III–V semiconductors). Such a simple substitution is not possible for acceptors, since the valence band structure of III–V semiconductors is much more complicated than the conduction band structure. The electronic band structure of several III–V semiconductors with zincblende structure was calculated by Chelikowsky and Cohen (1976). The band structure near the center of the Brillouin zone is schematically shown in ***Fig***. 15.7. The highest point of the valence band is located at $k = 0$. Without spin-orbit coupling this point would be sixfold degenerate with three dispersion relations and twofold spin degeneracy (Kohn, 1957a). The simplest way to understand this degeneracy is to consider the tight binding limit, in which the wave functions corresponding to the highest point go over into atomic 3p functions. The spin-orbit coupling lifts the degeneracy partially and leads to the situation shown in ***Fig***. 15.7. The top of the valence band remains fourfold degenerate at $k = 0$. The corresponding dispersion relations are called the heavy hole (hh) and light hole (lh) dispersion relations. The top of the twofold (spin) degenerate split-off

(so) band is at $k = 0$ at an energy E_{so} below the valence band maximum where E_{so} is the spin-orbit coupling energy. The top of the valence band corresponds to atomic $j = 3/2$ states (Kohn, 1957a), where j is the total angular momentum (orbit + spin), *i. e.* $j = l + s = 1 + 1/2 = 3/2$. The *inner quantum number j* is formally not necessary since it can be expressed by l and s. The introduction of j by Sommerfeld (1920) has historic reasons (Finkelnburg, 1958). The top of the valence band has Γ_8 symmetry. The split-off band corresponds to atomic $j = l - s = 1 - 1/2 = 1/2$ states which have Γ_7 symmetry.

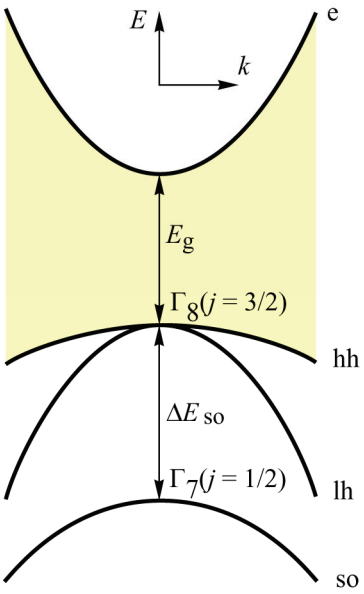


Fig. 15.7. Schematic electron, heavy hole (hh), light hole (lh), and split-off (so) dispersion relations near the center of the first Brillouin zone ($k = 0$).

For strong spin-orbit coupling, the split-off band is far removed from the top of the valence band, *i. e.* $\Delta E_{so} >> E_a$ where E_a is the acceptor binding energy. In this case, both the heavy hole and light hole band must be taken into account and the Hamiltonian is, therefore, a 4×4 matrix (Kohn, 1957a). In the limit of weak spin-orbit coupling, *i. e.* $\Delta E_{so} \approx E_a$, all three valence bands must be taken into account and the Hamiltonian is a 6×6 matrix.

Kohn (1957a) used a 6×6 Hamiltonian matrix to calculate acceptor energies in cubic semiconductors. The author used variational envelope wave functions for acceptors of the form

$$\psi_i(r) \quad = \quad A_i \, e^{-r/r_i} \tag{15.42}$$

where r_i is a variational parameter. Subsequently, Baldareschi and Lipari (1973) developed a now widely accepted model for shallow acceptor states in cubic semiconductors with degenerate valence bands. In their approach, the Hamiltonian is written as the sum of a spherical term and a cubic correction, thus pointing out the relevance of the spherical symmetry in the acceptor problem and the strong similarity to the case of atoms with spin-orbit interaction. Neglecting the cubic term, Hamiltonians with radial symmetry were obtained. Variational wave functions were used to calculate acceptor ionization energies.

In the limit of strong spin-orbit interaction, that is for spin-orbit splittings (ΔE_{so}) much larger than the acceptor energy (E_a), Baldareschi and Lipari (1973) calculated the effective Bohr radius. For an effective hole mass m_h^*, they obtained the effective Bohr radius

$$a_B^* = \frac{4\pi\varepsilon\hbar^2\gamma_1}{e^2 m_h^*}, \tag{15.43}$$

the effective Rydberg energy

$$E_{Ryd}^* = \frac{e^4 m_h^*}{2(4\pi\varepsilon\hbar)^2\gamma_1}, \tag{15.44}$$

and the acceptor ionization energy

$$E_a = E_{Ryd}^* f(\mu), \tag{15.45}$$

with

$$\mu = \frac{6\gamma_3 + 4\gamma_2}{5\gamma_1}. \tag{15.46}$$

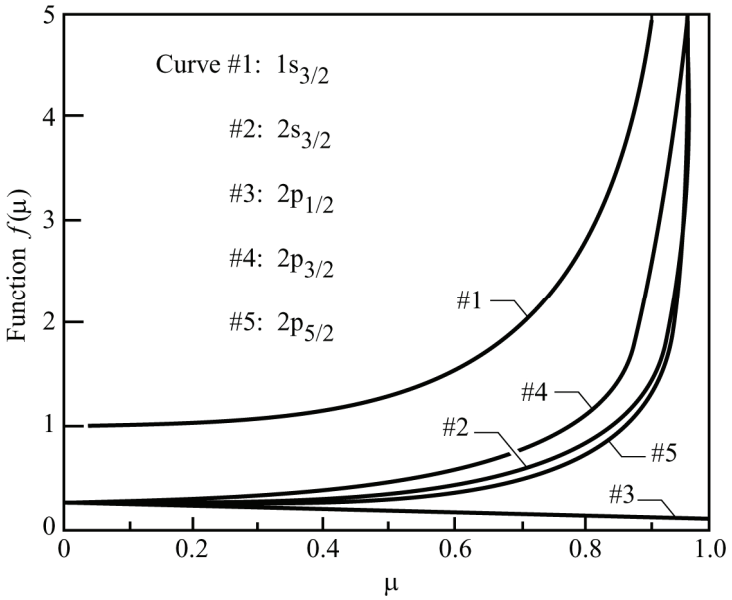


Fig. 15.8. Calculated function $f(\mu)$ versus μ in the limit of strong spin-orbit coupling for the spherical acceptor model (after Baldareschi and Lipari, 1973).

The parameters γ_1, γ_2, and γ_3 are the so-called Luttinger parameters which describe the hole dispersion relation near the center of the Brillouin zone (Luttinger, 1956). The function $f(\mu)$ relates the acceptor energy with the effective Rydberg energy (see Eq. (15.45)). The function $f(\mu)$ is shown in ***Fig***. 15.8. For the Luttinger parameters of III–V semiconductors, μ assumes values of $\mu \approx 0.6 - 0.9$ and the function $f(\mu)$ assumes values of $f(\mu) \approx 1.5 - 4$ for the ground state energy of acceptors. Baldareschi and Lipari (1973) obtained the numerical values of $f(\mu)$ using a variational approach and using spherical trial functions (Kohn, 1957a) for the acceptor wave functions (see Eq. 15.42). The Luttinger parameters, the variable μ, the effective Rydberg energy, and the ground state ($1s_{3/2}$) and excited state energies of hydrogenic acceptors in several III–V semiconductors were given by Luttinger (1956) and by Baldareschi and Lipari (1973). The calculated and experimental acceptor energies agree reasonably well.

In subsequent work, Baldareschi and Lipari (1974) investigated the contribution of cubic symmetry terms of the Hamiltonian to the spherical model for acceptor states. The effects of the cubic symmetry were studied using perturbation theory which allowed the authors to reproduce

all the details of acceptor spectra. The quantitative changes in acceptor energy caused by the cubic term are small, *i. e.* less than 1 meV for the III–V semiconductors.

The ***degeneracy*** of the acceptor ground states in III–V semiconductors is $g = 4$. Typical acceptor energies are much smaller than the spin-orbit splitting energy, *i. e.* $E_a << \Delta E_{so}$. The top of the valence band is fourfold degenerate due to heavy and light hole dispersion and due to twofold spin degeneracy. Since the acceptor wave functions are composed of valence band wave functions near the top of the band (Kohn, 1957a), the acceptor degeneracy is $g = 4$ as well.

15.4 Central cell corrections

The ionization energy of hydrogenic donors and acceptors as calculated from effective-mass theory does not depend on the chemical nature of the impurity atom. On the other hand, experimental values of the ionization energy do depend on the chemical nature especially for acceptors in III–V semiconductors. The difference in ionization energy between chemically different impurities is attributed to ***central cell potentials***. The central cell potential is assumed to be due to the chemical characteristics (*e. g.* electronegativity) of the impurity atom and thus the potential leads to a correction of the hydrogenic ionization energy. This correction is frequently referred to as ***chemical shift*** (Pantelides, 1975).

The total impurity potential is the sum of the Coulomb potential and the central cell potential V_{cc} and can be written as

$$V(r) = \frac{e}{4\pi\varepsilon r} + V_{cc}(r) \ . \tag{15.47}$$

The central cell potential is a short range potential and has a spatial extent of no more than the unit cell (central cell) of the host semiconductor. *Donor* wave functions are usually quite delocalized in III–V semiconductors. Therefore, the central cell corrections play a minor role for donors and their ionization energy is well described by the hydrogen model. In contrast, the *acceptor* Bohr radius is usually much smaller resulting in a significant central cell correction. This difference between donor and acceptor states is indeed observed experimentally, for example in GaAs. Several model potentials have been used for the central potential including a constant potential extending over the unit cell (Abarenkov and Heine, 1965), and δ-function-like potentials. Various models for central cell potentials were reviewed by Pantelides (1978), Stoneham (1975, 1986), and Altarelli and Bassani (1982). Finally, it is worthwhile to note that *isoelectronic* impurities lack the Coulomb term in Eq. (15.47). For isoelectronic impurities, the central cell potential is the only potential that can bind electrons (Thomas and Hopfield, 1966).

A simple model explaining the chemical shift of impurity ionization energies was proposed by Phillips (1970a, 1970b). The model is based on the local strain around the impurity atom. The strain is caused by the mismatch of the valence bonds of the impurity with valence bonds of the host lattice. Using this model, the chemical trend in donor ionization energies of Te, S, and Se in GaP were qualitatively explained (Phillips, 1970b). Phillips (1973) developed a second model in which the chemical shift in ionization energy is based on the difference in electronegativity, ΔX, between the impurity atom and the host lattice. The author showed that a large difference in electronegativity between impurity and the atom replaced by the impurity results in a large chemical shift. The chemical shift was assumed to be proportional to the heat of formation, *i. e.*

$$\Delta E \propto (\Delta X)^2 \tag{15.48}$$

where ΔE is the difference between the calculated effective mass impurity energy and the actual

impurity energy (*i. e.* ΔE is the chemical shift). Using Eq. (15.48), differences in chemical shifts of impurities occupying cation and anion sites in GaP were explained.

Note that chemical shifts are expected to be larger for impurity states with s-type symmetry as compared to states with p-type symmetry. The central cell potential is spatially restricted to the atomic vicinity of the impurity atom. In this region the amplitude of s-type wave functions is large while p-type wave functions have a node. Thus, perturbation theory predicts greater corrections for s-type symmetry states as compared to p-symmetry states.

15.5 Impurities associated with subsidiary minima

The conduction band structure of III–V semiconductors consists of three local minima, which occur at the *L*, Γ, and *X*-point of the Brillouin zone. The Γ minimum is located at the center of the Brillouin zone at $k = 0$, while the *L* and *X* minima occur at finite wave vectors. Donors, which by their very nature are associated with the conduction band, can form donor levels with all local minima of the conduction band. Impurity states associated with subsidiary minima of the conduction band were first analyzed by Bassani *et al.* (1969). The theoretical study revealed that the impurity states are generally formed from Bloch functions of many Brillouin zones and their respective contributions depend on the particular band structure and on the strength and nature of the impurity potential. The existence of several conduction band minima with large effective masses can increase the number of bound states as compared to the single valley hydrogenic model. It was further shown that resonant states in the continuum of one local minimum can be produced by impurity states associated with another minimum. This situation is shown in ***Fig***. 15.9 which depicts two donor levels associated with two conduction band minima and a resonant state in the continuum of the energetically lower minimum (Altarelli and Bassani, 1982).

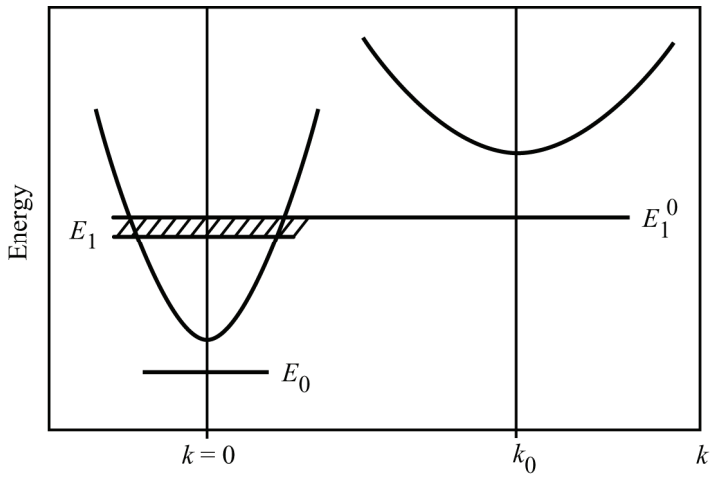


Fig. 15.9. Schematic representation of a bound state E_0 and a resonant state E_1^0 associated with the absolute minimum and a subsidiary minimum of the band structure, respectively. The resonant state E_1 (originating from the unperturbed state E_1) has a finite width due to its degeneracy with band states (after Altarelli and Bassani, 1982).

Donor levels associated with subsidiary conduction band minima were experimentally observed in III–V semiconductors. Adler (1969) used hydrostatic pressure to study the effect of donors on the electron transfer in n-type GaAs. Onton *et al.* (1972) directly observed a subsidiary conduction band minimum and its associated donor levels in InP by optical absorption measurements.

References

Abarenkov I. V. and Heine V. "The model potential for positive ions" *Philosophical Magazine* **12**, 529 (1965)

Adler, P. N. "Elevated Pressure Study of the Effect of Different Donors on Electron Transfer in *n*-GaAs" *Journal of Applied Physics* **40**, 3554 (1969)

Altarelli, M. and Bassani, F. in *Handbook on Semiconductors* **1**, edited by T. S. Moss and W. Paul (North Holland, Amsterdam, 1982)

Baldareschi A. and Lipari N. O. "Spherical Model of Shallow Acceptor States in Semiconductors" *Physical Review B***8**, 2697 (1973)

Baldareschi A. and Lipari N. O. "Cubic contributions to the spherical model of shallow acceptor states" *Physical Review B***9**, 1525 (1974)

Bassani F., Iadonisi G., and Preziosi B. "Band structure and impurity states" *Physical Review* **186,** 735 (1969)

Bethe H. (1942) unpublished

Bohm D. *Quantum Theory* (Prentice-Hall, Englewood Cliffs, 1951)

Borowitz S., *Fundamentals of Quantum Mechanics* (Benjamin, New York, 1967)

Chelikowsky, J. R. and Cohen, M. L. "Non-local pseudo-potential calculations for the electronic structure of eleven diamond and zincblende semiconductors**"** *Physical Review B***14**, 556 (1976)

Davydov A. S. *Quantum Mechanics* (Pergamon, Oxford, 1965)

Finkelnburg W. *Introduction to Atomic Physics* (in German) pg. 97 (Springer Verlag, Berlin, 1958)

Kohn W., in *Solid State Physics* **5**, edited by F. Seitz and D. Turnbull (Academic Press, New York (1957a)

Kohn W. "Effective Mass Theory in Solids from a Many-Particle Standpoint" *Physical Review*, **105**, 509 (1957b)

Luttinger J. M. "Quantum theory of cyclotron resonance in semiconductors: general theory" *Physical Review* **102**, 1030 (1956)

Madelung O. *Introduction to Solid-State Theory* (Springer Verlag, Berlin, 1978)

Merzbacher E. *Quantum Mechanics* (John Wiley and Sons, New York, 1970)

Onton A., Yacoby Y., Chicotka R. J. "Direct optical observation of the subsidiary X_{1c} conduction band and its donor levels in InP" *Physical Review Letters* **28**, 966 (1972)

Pantelides S. T. in *Advances in Solid State Physics Vol. 15,* edited by H. J. Queisser, p. 149 (Pergamon-Vieweg, Braunschweig, 1975)

Pantelides S. T. "The electronic structure of impurities and other point defects in semiconductors" *Reviews of Modern Physics* **50**, 797 (1978)

Philips J. C. "Dielectric Theory of Impurity Binding Energies: I. Group-V Donors in Si and Ge" *Physical Review B***1**, 1540 (1970a)

Philips J. C. "Dielectric theory of impurity binding energies. II. Donor and iso-electronic impurities in GaP" *Physical Review B***1**, 1545 (1970b)

Philips J. C., *Bonds and Bands in Semiconductors* (Academic Press, New York, 1973)

Saxon D. S., *Elementary Quantum Mechanics* (Holden-Day, San Francisco, 1968)

Sherwin C. W., *Introduction to Quantum Mechanics* (Henry Holt, New York, 1959)

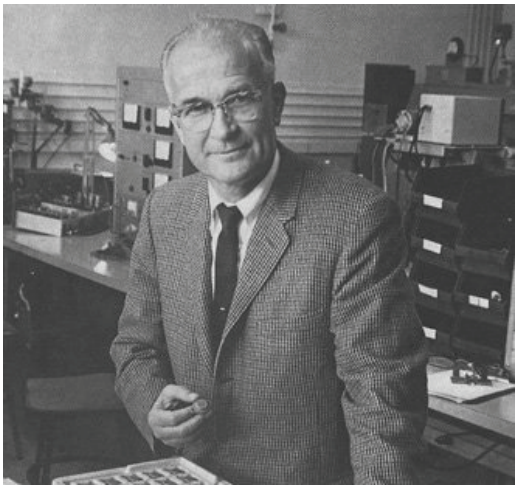
Sommerfeld A., see: Finkelnburg, W. (1958) *Introduction to Atomic Physics* (in German) (Springer Verlag, Berlin) pg. 97 (1920)

Stoneham A. M., *Theory of Defects in Solids* (Clarendon, Oxford, 1975)

Stoneham A. M., in Defects in Semiconductors, edited by H. J. Bardeleben (Trans. Tech., Switzerland, 1986)

Thomas D. G. and Hopfield J. J. "Isoelectronic traps due to nitrogen in gallium phosphide" *Physical Review* **150**, 680 (1966)

White H. E., *Introduction to Atomic Spectra* (McGraw-Hill, New York, 1934)



William B. Shockley (1910 – 1989)
Pioneered the physics of semiconductors and "father" of the transistor

16

High doping effects

16.1 Screening of impurity potentials

Variations of the electrostatic potential are reduced in magnitude by the spatial redistribution of free carriers. Variations in the band-edge potential can occur due to local doping concentration changes or local compositional changes of a semiconductor. An example of a potential fluctuation is shown in ***Fig***. 16.1. At some location, an excess positive donor charge causes a dip in the band-edge potential. The potential dip attracts electrons and results in a locally higher concentration of electrons at the potential dip. The potential generated by the negative charge of the excess electrons reduces the original fluctuation and smoothes the potential, *i. e.* free carriers *screen* the potential fluctuation. This is what screening of potential fluctuations by free carriers is all about. As an example of a potential fluctuation, consider the Coulomb potential of an impurity. As we recall, the impurity potential has a $1/r$ dependence. In the presence of free carriers the Coulomb potential is screened. The resulting potential is called the ***screened Coulomb potential*** which does *not* have the $1/r$ dependence of the unscreened Coulomb potential.

The potential variation is modeled in terms of the spatially non-uniform electrostatic potential $V(\mathbf{r})$, where $\mathbf{r} = (x, y, z)$ is the spatial coordinate. To find the energy levels in the perturbed potential $V(\mathbf{r})$, Schrödinger's and Poisson's equation must be solved simultaneously in order to find the free carrier distribution and the (screened) potential.

This procedure to calculate the screened potential is quite elaborate and it is usually possible to circumvent it, if the potential $V(\mathbf{r})$ is relatively smooth, *i. e.* it changes only little over the length of the electron wavelength. The electrons then 'see' only the potential at their own location. The total energy of the electron then follows the classical sum of potential and kinetic energy, *i. e.*

$$E = eV(\mathbf{r}) + \frac{\hbar^2 k^2}{2 m_e^*} . \qquad (16.1)$$

In other words, we assume that the spatial dimensions of the potential perturbation are so large that size quantization of the carrier system does not need to be considered.

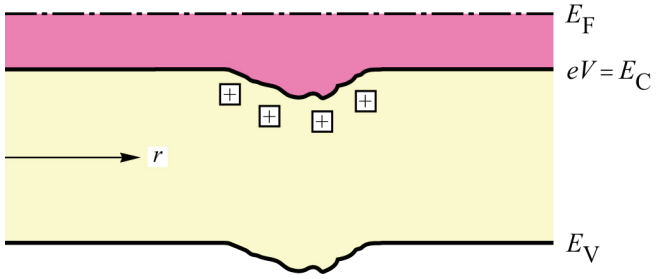


Fig. 16.1 Perturbation of the conduction band edge by an excessive number of positive charges. Electrons accumulate in the potential minimum and screen (*i. e.* compensate) the excess positive charge.

The electron concentration in the potential perturbation shown in ***Fig***. 16.1 is next calculated using classical quantization, *i. e.* using the semi-classical density of states. The conduction band edge sufficiently far away from the perturbation has a potential V_0 and an electron concentration of n_0. The electron concentration at the potential perturbation is given by

$$n(\mathbf{r}) \;=\; \int_{eV(\mathbf{r})}^{\infty} f_{\mathrm{FD}}(E)\,\rho_{\mathrm{DOS}}(E)\,\mathrm{d}E \tag{16.2}$$

where $f_{\mathrm{FD}}(E)$ and $\rho_{\mathrm{DOS}}(E)$ are the Fermi–Dirac distribution and the density of states, respectively. In the following, we differentiate between a degenerate and a non-degenerate electron gas. The results obtained for the two cases are, as will be seen, quite different. For a *non-degenerate* electron gas, Boltzmann statistics is employed and the local electron concentration is obtained as

$$n(\mathbf{r}) \;=\; \frac{1}{\sqrt{2}}\left(\frac{m_{\mathrm{e}}^{*}\,kT}{\pi\,\hbar^2}\right)^{3/2} \exp\frac{E_{\mathrm{F}} - e\,V(\mathbf{r})}{kT} \tag{16.3}$$

where E_{F} is the Fermi level. For a *degenerate* electron gas, Fermi–Dirac statistics must be employed and the electron concentration in the limit of *extreme degeneracy* [$E_{\mathrm{F}} - eV(\mathbf{r}) >> kT$] is given by

$$n(\mathbf{r}) \;=\; \frac{(2\,m_{\mathrm{e}}^{*})^{3/2}}{3\pi^2\,\hbar^3}\left(E_{\mathrm{F}} \;-\; e\,V(\mathbf{r})\right)^{3/2} . \tag{16.4}$$

If the potential perturbation is a 'dip' in the conduction band then $E_{\mathrm{F}} - eV(\mathbf{r}) > E_{\mathrm{F}} - eV_0$ and electrons accumulate in the dip. If the perturbation is a 'bump' in the conduction band then $E_{\mathrm{F}} - eV(\mathbf{r}) < E_{\mathrm{F}} - eV_0$ and electrons deplete at the location of the perturbation. In the case of a 'dip', the potential generated by the excess electrons reduces the magnitude of the dip, *i. e.* smoothes the potential. The resulting potential is obtained from Poisson's equation, which relates the charge density and the potential $V(\mathbf{r})$ according to

$$\nabla^2\,V(\mathbf{r}) \;=\; -\frac{e}{\varepsilon}\left[\xi(\mathbf{r}) \;-\; n(\mathbf{r}) \;-\; n_0\right] \tag{16.5}$$

where $e\,\xi(\mathbf{r})$ is the concentration of fixed (positive) charge, $\nabla = (\partial/\partial x,\ \partial/\partial y,\ \partial/\partial z)$ is the Nabla operator, and $\nabla^2 = (\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2)$. The concentration $\xi(\mathbf{r})$ is homogeneous except at the location of interest. The average concentration of electrons is n_0, while $n(\mathbf{r})$ is the deficiency or excess of charge which depends on $\mathbf{r}$. Equation (16.5) is called a quasi-classical equation of the Thomas–Fermi type. Since $n(\mathbf{r})$ depends in a non-linear manner on $V(\mathbf{r})$, the differential equation is non-linear. It is the basic equation of non-linear screening theory. Unfortunately, the equation has no general solution. However, in the limit of small variations $n(\mathbf{r})$, it is possible to linearize the differential equation. Suppose the potential fluctuations of $V(\mathbf{r})$ are small as compared to $E_{\mathrm{F}} - eV(\mathbf{r})$. Then Eqs. (16.3) and (16.4) can be linearized, *i. e.*

$$\boxed{n(\mathbf{r}) \;=\; n_0 \;+\; \left.\frac{\mathrm{d}n}{\mathrm{d}V(\mathbf{r})}\right|_{n=n_0} V(\mathbf{r})} \tag{16.6}$$

which greatly simplifies screening theory. The linearization of $n(\mathbf{r})$, *i. e.* our restriction to *small* potential fluctuations is the basis of ***linear screening theory***. The linearization of Eq. (16.6) allows one to write the Poisson equation (Eq. 16.5) as

$$\nabla^2 V(\mathbf{r}) = \frac{V(\mathbf{r})}{r_s^2} - \frac{e\,\xi(\mathbf{r})}{\varepsilon} \tag{16.7a}$$

where

$$r_\mathrm{s} = \left[-\left(\frac{\mathrm{d}n}{\mathrm{d}V} \right)_{V=0} \frac{e}{\varepsilon} \right]^{-1/2} \tag{16.7b}$$

is called the screening radius. The screening radius is usually referred to as the ***Debye screening radius*** and the ***Thomas–Fermi screening radius*** for non-degenerate and degenerate electron systems, respectively. Using Eq. (16.7b) and non-degenerate statistics (Boltzmann) and degenerate statistics (Fermi–Dirac) for an isotropic and parabolic conduction band, the screening radii are, respectively, obtained as:

Debye screening radius

$$\boxed{r_\mathrm{D} = \sqrt{\varepsilon\, kT/(e^2\, n)}} \tag{16.8}$$

Thomas–Fermi screening radius

$$\boxed{r_\mathrm{TF} = \pi^{2/3} \sqrt{\frac{\varepsilon\, \hbar^2}{e^2\, m^*\, (3n)^{1/3}}}} \tag{16.9}$$

Note the different functional dependences of the Debye and the Thomas–Fermi screening radius on temperature and free carrier concentration. While r_D depends on temperature, r_TF is temperature-independent. Furthermore, the Thomas–Fermi screening radius depends very weakly on the electron concentration, *i. e.* $n^{-1/6}$. Now the analytic solution of the screened potential can be obtained by integration of Eq. (16.7a) and the solution is given by

$$V(\mathbf{r}) = \int_r K(\mathbf{r}-\mathbf{r}')\ \xi(\mathbf{r}')\ \mathrm{d}\mathbf{r}' \tag{16.10}$$

with $K(r) = [e/(4\pi\varepsilon r)]\, \mathrm{e}^{-r/r_\mathrm{s}}$.

We now calculate the screened potential of a single ionized impurity. The charge distribution of such a single impurity located at the origin of the coordinate system is $\xi(\mathbf{r}) = \delta(\mathbf{r})$. Insertion into Eq. (16.10) yields the ***screened Coulomb potential***

$$\boxed{V(r) = \frac{e}{4\pi\varepsilon r}\, e^{-r/r_\mathrm{s}}} \tag{16.11}$$

where the screening radius, r_s, can be either the Debye or the Thomas–Fermi radius depending on the degeneracy of the screening electron gas. The screened Coulomb potential (Debye and Hückel, 1923) is also called the ***Yukawa potential*** in analogy to a potential in meson theory. The screened Coulomb potential is modified as compared to the unscreened Coulomb potential by the factor $\exp(-r/r_\mathrm{s})$. Note that for $r \to 0$ the screened and the unscreened Coulomb potential become identical. For $r \to \infty$ the screened and unscreened Coulomb potential are strongly diverging due to the $\exp(-r/r_\mathrm{s})$ factor. The screened and the unscreened Coulomb potential are

illustrated in ***Fig***. 16.2 for an impurity located at $r = 0$. The unscreened potential is a long range potential while the screened one has a shorter spatial range.

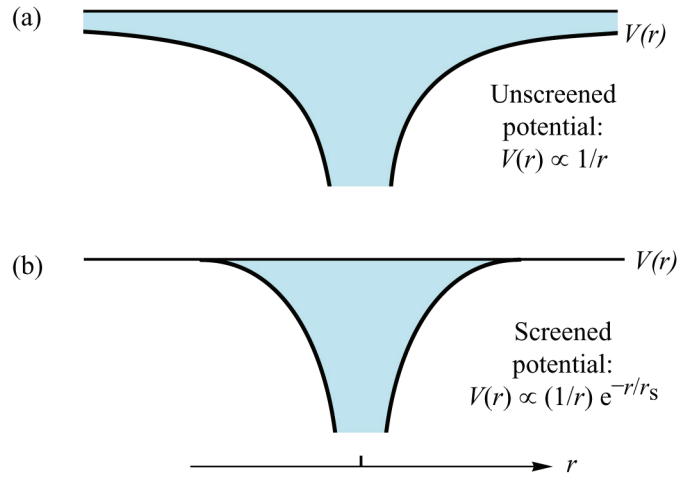


Fig. 16.2 (a) Unscreened and (b) screened Coulomb potentials. The unscreened and screened Coulomb potentials coincide for $r \to 0$ but strongly deviate for large radii.

There are several major results which are inferred from linear screening theory which are summarized in the following. *First*, the screening of potential perturbations leads to an exponential decay of the perturbing potential with distance, *i. e.* $\exp(-r / r_s)$, where r_s is the screening radius. *Second*, in the case of several potential perturbations, each perturbation is screened separately. According to Eq. (16.10), a potential fluctuation can be expressed as a superposition of (unscreened) Coulomb potentials. The screened potential is then given by the superposition of the *screened Coulomb potentials. The applicability of the superposition principle is a consequence of the linearity* of the screening theory. *Third*, the screening radius r_s does not depend on the *magnitude* of the potential perturbation to be screened. Instead the screening radius depends on the properties of the screening electron (hole) gas only. The independence of the screening radius on the magnitude of the potential fluctuation is again due to the linearization of screening theory.

The linear screening theory becomes invalid for very large potential fluctuations. The linearization of screening theory given in Eq. (16.6) is based on the assumption of *small* fluctuations. As an example of a large potential perturbation, consider a metal–semiconductor junction (Schottky barrier). Such junctions can induce perturbations of *e. g.* 1 eV (Schottky barrier height). For such large potential perturbations ***non-linear screening theory*** must be employed to obtain realistic results. Linear screening theory cannot be applied to Schottky barriers. The dependence of the depletion region thickness (*i. e.* screening length) on the magnitude of the Schottky barrier height of a metal–semiconductor interface is a result of non-linear screening theory.

The nature of screening changes drastically, if the screening carriers are confined to a two-dimensional plane. This situation is referred to as ***two-dimensional screening***. Ando *et al.* (1982) considered a Coulomb charge located at the cylindrical coordinates $r = 0$ and $z = z_0$ which is screened by electrons confined to the plane $z = 0$. The screened Coulomb potential in the electron plane ($z = 0$) is given by

$$V(r, z = 0) \quad = \quad \int_0^\infty q\, A(q)\, J_0(qr)\, \mathrm{d}q \tag{16.12}$$

where J_0 is the Bessel function of zero order. The constant $A(q)$ is

$$A(q) \quad = \quad \frac{e}{4\pi\varepsilon} \frac{\mathrm{e}^{qz_0}}{q + 1/r_s^{2D}} \tag{16.13}$$

where r_s^{2D} is the two-dimensional screening radius

$$\boxed{r_s^{2D} \quad = \quad \frac{2\varepsilon}{e^2} \frac{\pi\hbar^2}{m^*}} \tag{16.14}$$

Note that the two-dimensional screening radius depends neither on temperature, nor on the sheet charge density. This characteristic is a result of the two-dimensional density of states which is constant and does not depend on energy. For large values of r, where $r / r_s^{2D} >> 0$ the asymptotic form of the potential seen by the electrons is (Stern, 1967)

$$V(r, z=0) \quad \approx \quad \frac{e\left(1 + z_0 / r_s^{2D}\right)}{4\pi v \varepsilon r^3 / (r_s^{2D})^2} \ . \tag{16.15}$$

This inverse-cube dependence of the potential on distance is much weaker than the exponential decay found in the three-dimensional case, and is one of the principal qualitative differences between two-dimensional and three-dimensional screening (Ando *et al.*, 1982).

16.2 The Mott transition

At high doping concentrations, many characteristics of semiconductors change. The changes are due to either the high concentration of impurities or due to the high free carrier concentration accompanying the high doping concentrations. Among the characteristics changed at high doping concentrations are the impurity ionization energy, the fundamental absorption edge, the density of states in the vicinity of the band edges, and the energy of the fundamental gap. Effects causing these changes are the Mott transition, the Burstein–Moss shift, band tailing effects, and bandgap renormalization.

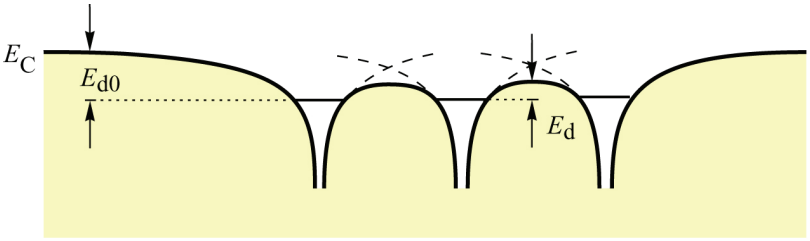


Fig 16.3. Conduction band edge with three donor potentials. For high donor concentrations, the Coulomb potentials overlap and the ionization energy E_{d0} is reduced to E_d.

The Mott transition refers to an ***insulator-to-metal*** transition occurring in semiconductors at high doping concentrations. Consider an n-type semiconductor with low doping concentration. At low temperatures ($T \to 0$), shallow impurities are neutral, *i. e.* electrons occupy the ground state of the donor impurities. All continuum states in the conduction band are unoccupied. In this case, the semiconductor has the properties of an *insulator*. As the doping concentration increases, the Coulomb potentials of impurities overlap as schematically shown in ***Fig***. 16.3. As a result of the overlapping impurity potentials, electrons can transfer more easily from one donor to another donor. Electrons transfer from one donor state to a state of an adjacent donor by either

tunneling or by *thermal emission* over the barrier. The probability of both processes increases with decreasing donor separation. In other words, the activation energy for electron transport is reduced. In the extreme case, the activation energy approaches zero, *i. e.* the conductivity remains finite even for $T \to 0$. The semiconductor then has *metal*-like properties.

Screening of impurity potentials is a second contribution to the reduction of the impurity ionization energy. Impurity potentials are effectively screened at high free carrier concentrations. Screened potentials are less capable of binding electrons. Thus, the effective ionization energy decreases due to screening.

The insulator-to-metal transition occurs at the impurity concentration at which the distance between impurities becomes comparable to the Bohr radius. If donors with concentration N were to occupy sites of a simple cubic lattice, their separation would be $N^{-1/3}$. The Mott transition would then occur at a concentration

$$2\, a_B^* \quad = \quad N_{crit}^{-1/3} \tag{16.16a}$$

where a_B^* is the effective Bohr radius and N_{crit} is called the critical concentration. However, Eq. (16.16a) does not give the correct result, because impurities are distributed *randomly* in semiconductors. Using a poissonian distribution of impurities, one can show that the Bohr orbital of an impurity is likely to overlap with the orbitals of one, two, or three neighboring impurities if

$$2a_B^* \quad = \quad \frac{3}{2\pi}\, N_{crit}^{-1/3} \; . \tag{16.16b}$$

Rearrangement of the equation yields the ***Mott criterion***

$$\boxed{a_B^*\, N_{crit}^{1/3} \quad \approx \quad 0.24} \tag{16.16c}$$

As an example, we calculate the critical concentration of donors in GaAs with an effective Bohr radius of $a_B^* = 103$ Å. Equation (16.16c) yields a critical density of $N_{crit} = 1.2 \times 10^{16}$ cm^{-3}.

The Mott criterion can be also obtained by a fundamentally different approach, *i. e.* by considering the binding of electrons to screened Coulomb potentials (see, for example, Mott, 1990). At low concentrations, electrons are bound to the (essentially) unscreened Coulomb potentials of the impurities. At higher free carrier concentrations, screening becomes relevant and Coulomb potentials must be replaced by screened Coulomb (Yukawa) potentials. The binding energy of screened Coulomb potentials is smaller than the binding energy of Coulomb potentials. Furthermore, the binding energy decreases with increasing screening. The insulator-to-metal transition occurs, if the binding energy of electrons bound to screened Coulomb potentials becomes zero.

To solve this problem quantitatively, we use the Coulomb potential $V(r) = [\, e / (4\pi\varepsilon r)\,]$ $\exp(-r / r_{TF})$, where r_{TF} is the screening radius. Furthermore, we use the variational wave function $\psi(r) = c \exp(-r/\alpha)$, where c is a normalization constant and α is the variational parameter (Flügge, 1971). The binding energy, E, and the spatial extent of the wave function, α, can be obtained by the variational method. From the condition $E = 0$ one obtains the variational parameter $\alpha = 2\, a_B^*$ and $r_{TF} / a_B^* = 1$. Using Eq. (16.9) for r_{TF} and the equation for the effective Bohr radius of hydrogenic impurities, a_B^*, yields

$$\boxed{a_B^*\, N_{crit}^{1/3} \quad = \quad \frac{1}{4}\left(\frac{\pi}{3}\right)^{1/3} \quad \approx \quad 0.25} \tag{16.17}$$

which is similar to Eq. (16.16c).

There is yet another approach to calculate the Mott density. The calculation is based on the reduction of the bandgap energy due to many-body effects. Haug and Schmitt-Rink (1984) showed that the energy level of an exciton merges with the conduction band due to the lowering of the conduction band edge at high concentrations (see Sect. on *bandgap renormalization*). The authors further showed that the energy of the exciton is remarkably constant with carrier concentration due to the charge neutrality of the exciton. The critical concentration, estimated from the bandgap reduction due to many-body effects (see Sect. on *many-body effects*), agrees well with the Mott criterion of Eq. (16.17). Even though the result of Haug and Schmitt-Rink (1984) was obtained for excitons, it also applies to neutral donors which can be thought of as excitons with an infinitely heavy hole mass.

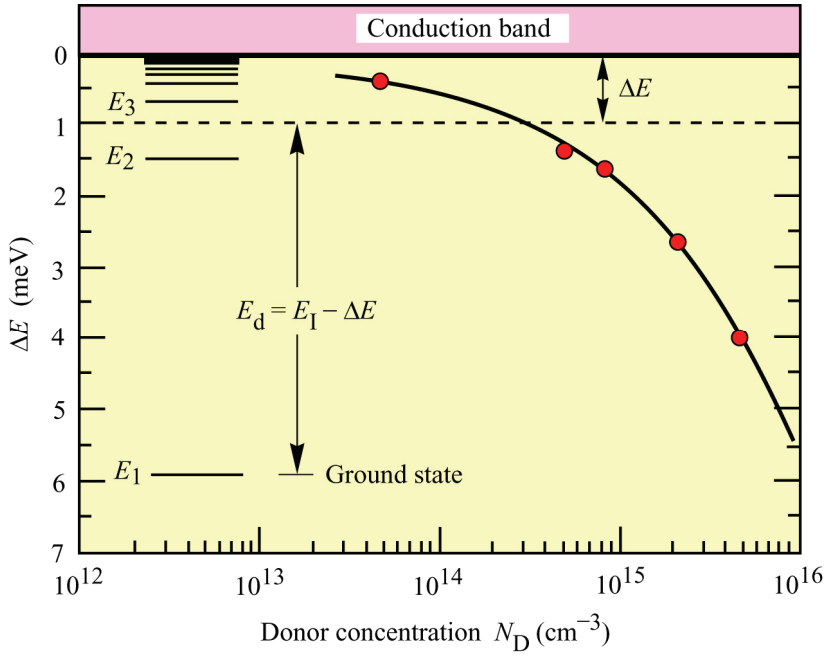


Fig. 16.4. Hydrogenic donor levels in GaAs, E_n, and measured donor ionization energy E_d as a function of doping concentration (Stillman *et al*., 1982).

The insulator-to-metal transition does not occur abruptly at the critical concentration. Instead the transition is gradually evolving with increasing impurity concentration. Quantitatively, the gradual nature of the Mott transition can be expressed in terms of a continuously changing impurity activation energy. Experimental donor activation energies have been described by the equation (Debye and Conwell, 1954)

$$E_d = E_{d0}\left[1 - \left(N_D / N_{crit}\right)^{1/3}\right] \tag{16.18}$$

where E_{d0} is the donor activation energy for $N_D << N_{crit}$. The reduction in donor ionization energy is thus proportional to the distance between the donor atoms. The ionization energy of donors in GaAs as a function of the donor concentration is shown in ***Fig***. 16.4 (Stillman *et al.*, 1982). The effective ionization energy is measured from the impurity level to the quasi-continuum. The effective ionization energy approaches zero at the critical density which is $N_{crit} \approx 10^{16}$ cm^{-3} for donors in GaAs.

16.3 The Burstein–Moss shift

The shift of the absorption edge to higher energies occurring at high doping concentrations is referred to as the ***Burstein–Moss effect*** or ***shift*** (Burstein, 1954; Moss, 1961). This shift is also called ***band filling*** or ***phase space filling***. The up-shift of the absorption edge is related to band filling which is schematically shown in ***Fig***. 16.5 for n-type doping. The conduction band becomes significantly filled at high doping concentrations due to the finite density of states. Due to band filling, absorption transitions cannot occur from the top of the valence band to the bottom of the conduction band. As a result, the fundamental edge of absorption transitions shifts from $E_C - E_V = E_g$ for undoped semiconductors, to $E_F - E_V > E_g$ in heavily doped n-type semiconductors. This shift was first observed in n-type InSb (Moss, 1961) and has since been used to make semiconductors transparent in the near-band-edge region (Verie, 1967; Dapkus *et al.*, 1969, Deppe *et al.*, 1990). It should be pointed out that the Burstein–Moss shift competes with the formation of impurity bands (see Sect. on *impurity bands*) and band tails (see Sect. on *band tails*). As will be seen, the Burstein–Moss shift dominates in semiconductors with light effective carrier mass.

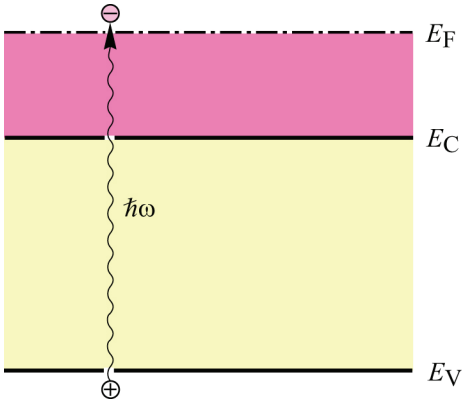


Fig. 16.5. Schematic illustration of conduction band filling due to heavy n-type doping. The absorption edge occurs at an energy $E_F - E_C$ which can be significantly larger than the bandgap energy $E_C - E_V$.

Quantitatively, the Burstein–Moss shift can be calculated from the filling of the conduction or valence band in n-type and p-type semiconductors, respectively. Assuming a degenerately doped n-type semiconductor, the band filling for a single-valley, isotropic, and parabolic band is in the limit of extreme degeneracy given by

$$E_F - E_C = \frac{\hbar^2}{2m_e^*}(3\pi^2 n)^{2/3} \tag{16.19}$$

where n is the free carrier density and m_e^* is the effective carrier mass. The absorption edge occurs then at the energy

$$E = E_g + (E_F - E_C) \; . \tag{16.20}$$

Note that the Burstein–Moss shift is inversely proportional to the effective mass. This explains the fact that the Burstein–Moss effect is more prominent in semiconductors with light carrier masses. For example, the shift clearly manifests itself in n-type GaAs ($m_e^* = 0.067\ m_0$) but not in p-type GaAs ($m_{hh}^* = 0.45\ m_0$) as illustrated in ***Figs***. 16.6 and 16.7.

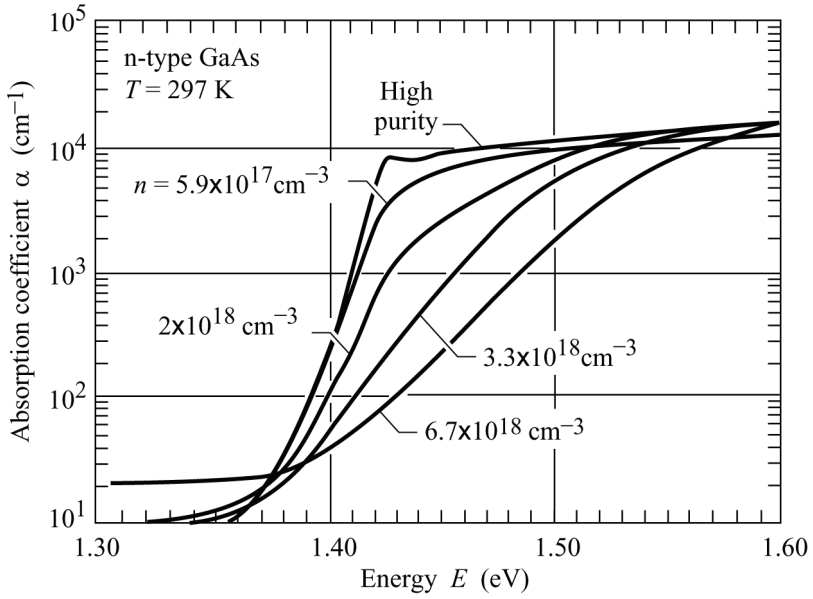


Fig. 16.6 Absorption coefficient in n-type GaAs at T = 297 K for different doping concentrations. The absorption edge shifts to higher energies with increasing doping concentration (after Casey *et al.*, 1975).

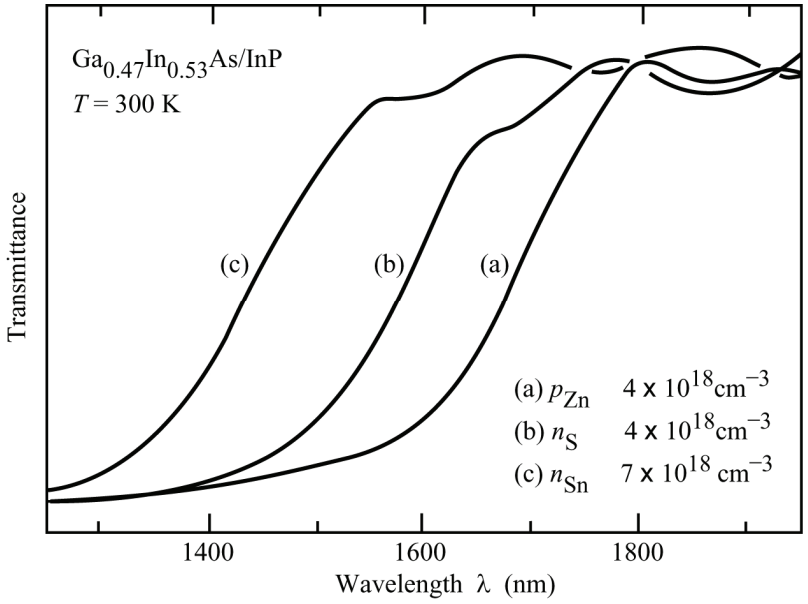


Fig. 16.7 Optical transmittance versus wavelength for p- and n-type $Ga_{0.47}In_{0.53}As$ epitaxial layers grown on InP. The layers are doped with Zn, S, and Sn. Near-band-edge absorption decreases as the n-type doping level increases (after Deppe *et al.*, 1990).

For the sake of completeness, we consider the Burstein–Moss effect in *two-dimensional systems*. In such systems, free carriers are confined to a thin sheet. The confinement can be achieved in terms of quantum well structures or at the heterojunctions of two semiconductors. The band filling of a semiconductor with a single, parabolic subband is, in the high doping limit, given by

$$E_F - E_0 = \frac{\pi \hbar^2}{m_e^*} n^{2D} \tag{16.21}$$

where E_0 is the bottom of the subband and n^{2D} is the two-dimensional carrier concentration per cm^2. The energy given by Eq. (16.21) represents the shift with respect to the absorption edge in

an undoped two-dimensional semiconductor.

The Burstein–Moss shift is illustrated in ***Fig***. 16.6 for n-type GaAs (Casey *et al.*, 1975). The graph shows the absorption coefficient as a function of photon energy for n-type doping concentrations up to 6.7×10^{18} cm^{-3}. As the carrier concentration exceeds 6×10^{17} cm^{-3}, the Burstein–Moss shift due to the filling of the conduction band begins to have a significant effect. This shift of the absorption coefficient to higher energies is readily shown in ***Fig***. 16.6 for samples with $5.9 \times 10^{17} < n < 6.7 \times 10^{18}$ cm^{-3}. These curves tend to converge at 1.6 eV to within ± 10 % of the α of the high-purity sample. For the highly doped sample, absorption of $\alpha = 20$ cm^{-1} occurs for $E < 1.38$ eV, which is probably related to band tails (see Sect. on *band tails*). Note that the absorption spectra of n-type GaAs are in stark contrast to absorption in p-type GaAs (Casey *et al.*, 1975). The difference is due to the much heavier mass of holes as compared to electrons, which makes the Burstein–Moss shift less important in p-type GaAs.

Transmission spectra of a $Ga_{0.47}In_{0.53}As$ lattice matched to InP are shown in ***Fig***. 16.7 for various doping concentrations (Deppe *et al.*, 1990). For a p-type Zn doping level of $N_{Zn} \approx 4 \times 10^{18}$ cm^{-3}, no appreciable shift of the absorption edge is observed as compared to undoped bulk $Ga_{0.47}In_{0.53}As$ ($\lambda_{GaInAs,\ 300\ K} \approx 1.65$ μm). However, a significant shift is observed for n-type $Ga_{0.47}In_{0.53}As$ at a doping level of 4×10^{18} cm^{-3} and an even larger shift at a doping level of 7×10^{18} cm^{-3}. At the highest n-type doping level, the semiconductor is virtually transparent at a wavelength of 1.55 μm. The highly n-type doped $Ga_{0.47}In_{0.53}As$ was used for quarter-wave reflectors operating at 1.55 μm, which require optical transparency at that wavelength (Deppe *et al.*, 1990).

16.4 Impurity bands

At impurity concentrations well below the critical Mott concentration, impurities can be considered as isolated, non-interacting entities. As the concentration increases but is still well below the Mott concentration, impurities begin to *interact*. Carrier transport at low temperatures occurs via thermally assisted tunneling between impurity states. This transport process is called ***hopping conduction***. At still higher impurity concentrations but below the critical Mott concentration, overlapping impurity states form an ***impurity band***. At low temperatures, carriers can propagate within the impurity band without entering the conduction band. This transport process is known as ***impurity band conduction***. This section summarizes the elementary characteristics of impurity-assisted conduction mechanisms for concentrations below the Mott concentration. Extensive reviews of the topic were given by Mott (1987, 1990) and Shklovskii and Efros (1984).

Consider a semiconductor with a donor density well below the critical Mott transition density. Upon cooling the semiconductor to low temperature, the conductivity is expected to decrease as free electrons freeze out onto localized donor states. For $kT << E_d$ the conductivity of an n-type semiconductor is expected to become vanishingly small. Experimentally, zero conductivity is not observed in semiconductors containing a net concentration of shallow impurities. Instead, the temperature dependence of the conductivity is less drastic than expected from free carrier freeze-out. The conductivity in this regime is not given by electrons excited to the conduction band but rather by electrons hopping from neutral donors to ionized donors. The conductivity is referred to as ***hopping conductivity*** and can be described by the general hopping conductivity formula

$$\sigma_{hop} = \left(\alpha / T^{\beta}\right) \exp\left(-\frac{E_{hop}}{kT}\right)^{\gamma} \qquad (16.22)$$

where α and β are constants, E_{hop} is the thermal activation energy for the hopping process and γ determines the functional dependence of the exponential factor. For simplicity, the factor γ is frequently assumed to be unity. Austin and Mott (1969) showed for gaussian localization, in which transport occurs via tunneling to remote but energetically similar donors, that the value of $\gamma = 1/4$. Efros and Shklovskii (1975) showed that $\gamma = 1/2$, if Coulomb interaction between adjacent donors determines the hopping transport between adjacent donors. In this case the thermal activation energy for the hopping process is given by

$$E_{hop} \approx \alpha \frac{e^2}{4\pi\varepsilon} \left(\frac{4\pi}{3} N_D \right)^{1/3} \tag{16.23}$$

where the factor α has a numerical value of about 0.60. The qualitative physical explanation of the hopping conduction process is as follows. Consider a lightly compensated n-type semiconductor at low temperatures. Most of the donors are neutral, some donors ionized due to the slight compensation. When the donor impurities are closely spaced, their energy levels split. Electrons can tunnel from a donor state to an empty state of an adjacent donor. A so-called ***Coulomb gap*** develops between filled donor states and empty donor states. The Coulomb gap is caused by the long-range Coulomb interaction of localized electrons (Pollak and Knotek, 1974; Efros and Shklovskii, 1975), and occurs at the Fermi level. The tunneling from filled donor states to adjacent empty donor states therefore requires a small thermal activation energy given by Eq. (16.23). The activation energy can be interpreted as the energy from the Fermi level to the energy of the maximum of the density of empty state distribution (Böer, 1990). Typically the activation energy is much smaller than the donor ionization energy.

The regime of hopping conduction is schematically illustrated in ***Fig***. 16.8 which shows the dispersion relations and donor impurity states at different doping densities. At low doping density, states of adjacent impurities do not interact ($N << N_{crit}$) and impurity ground states are discrete and energetically well-defined levels. As the doping concentration increases, states of adjacent impurities interact, split, and finally form an impurity band ($N \leq N_{crit}$). At even higher doping concentrations, the impurity band widens and merges with the continuum band ($N \geq N_{crit}$).

It is instructive to consider the effect of *ordered* and *random* impurity distribution on the formation of an impurity band, which is shown in ***Fig***. 16.8(a) and (b), respectively. In the case of an *ordered* impurity distribution, the potentials of the impurities are strictly periodic, similar to the periodic potential assumed in the Kronig–Penney model. As a result, the impurity band has a well-defined width and well-defined edges, as shown in ***Fig***. 16.8(a). However, the case of *random* impurity distribution in semiconductors is a more realistic assumption (Shockley, 1961). For a random impurity distribution, the impurity band does not have well-defined band edges but the impurity states will tail into the forbidden gap, as shown in ***Fig***. 16.8(b). Tail states of impurities occurring at high doping density are discussed in a subsequent Section. ***Figure*** 16.8(c) schematically shows hopping conduction and impurity band conduction.

Impurity bands are formed at sufficiently high doping concentrations. The impurity band is formed by the quasi-periodic Coulomb potentials of the impurities. Since the impurity potentials are not strictly periodic, a periodic band calculation (*e. g*. Kronig–Penney) cannot be used. The hydrogenic potentials of donors can be used to form a Hubbard band (Hubbard, 1963). In the Hubbard model, the center of the impurity band coincides with the energy level of the unperturbed impurity state. If the donors forming the impurity band are partially compensated by acceptors, the band is only partially filled and impurity band conduction can occur within the impurity band, *i. e.* electrons need not occupy conduction band states for carrier transport (Adler,

1982). The impurity band width, ΔE_D, is approximately given by the overlap integral between donors separated by the average distance $N_D^{-1/3}$. The band width is approximately equal to the interaction energy

$$\Delta E_D \approx \frac{e^2}{4\pi\varepsilon N_D^{-1/3}} . \tag{16.24}$$

As an example, we consider GaAs with a donor concentration of 1×10^{16} cm^{-3} and $\varepsilon_r = 13.1$ and obtain $\Delta E_D \approx 2.4$ meV.

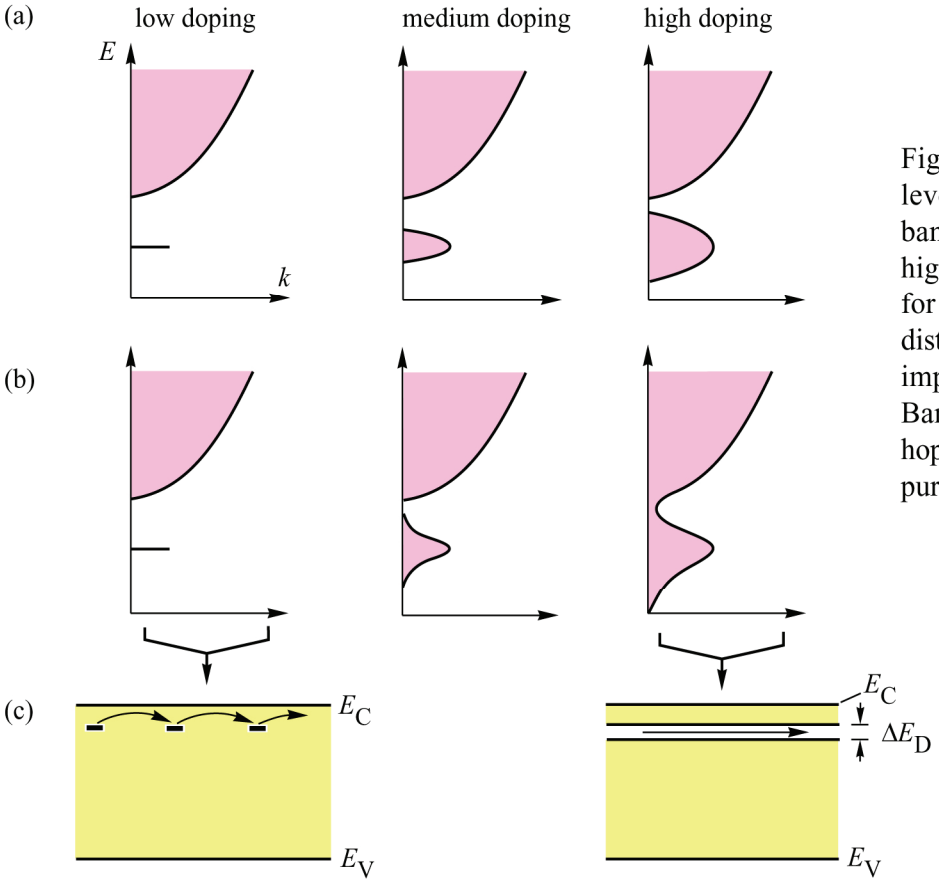


Fig. 16.8. Donor impurity level and donor impurity band at low, medium and high doping concentrations for (a) an ordered impurity distribution and (b) a random impurity distribution. (c) Band diagram illustrating hopping conduction and impurity band conduction.

The mobility associated with impurity band conduction are very low, typically < 1 cm^2/Vs. The low mobility is due to the heavy dispersion mass. In order to qualitatively relate the low impurity band mobility with the narrow width of the impurity band, let us recall some of the basic results of one-dimensional band theory. The dispersion mass of a parabolic band is given by

$$m^* = \frac{\hbar^2}{\mathrm{d}^2E/\mathrm{d}k^2} . \tag{16.25}$$

In the simple Kronig–Penney model, the dispersion (*i. e.* effective) mass is given by

$$m^* = 2\hbar^2 / \left(z_p^2 \, \Delta E_D \right) \tag{16.26}$$

where z_p is the period of the periodic potential and ΔE_D is the width of the band. Although the result of the one-dimensional Kronig–Penney model cannot be rigorously applied to an impurity band, the model provides a qualitative understanding of the functional dependences. It is therefore reasonable to conclude that a heavy effective mass is associated with transport of carriers in a narrow impurity band. Since the mobility and the effective mass are related according to the Drude model by

$$\mu = \frac{e\tau}{m^*}, \tag{16.27}$$

a low mobility results for carrier transport in a partially filled impurity band.

16.5 Band tails

The random distribution of charged impurities results in potential fluctuations of the band edges. In an *undoped* semiconductor such potential fluctuations are absent and the band edges are well defined. In a highly doped semiconductor the potential fluctuations cause the band edges to vary spatially. This situation is schematically shown in ***Fig***. 16.9. States with energy below the unperturbed conduction band edge or above the unperturbed valence band edge are called ***tail states***. The tail states significantly change the density of states in the vicinity of the band edge.

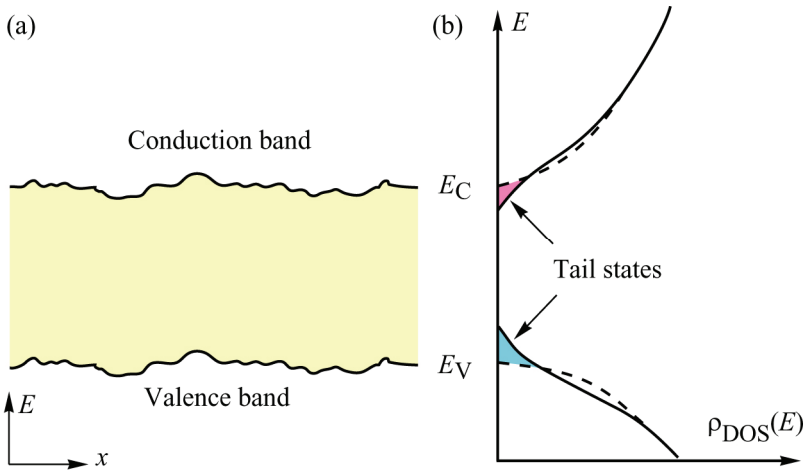


Fig. 16.9. (a) Spatially fluctuating band edges caused by random distribution of impurities. (b) Resulting densities of states in the conduction and valence band with tail states extending into the forbidden gap. The dashed lines show the parabolic densities of states in undoped semiconductors.

The magnitude of band-edge energy fluctuations caused by the random distribution of charged donors and acceptors was first calculated by Kane (1963). For an ionized donor concentration of N_D and an ionized acceptor concentration of N_A, the root-mean-square (energy) fluctuation of the band edges is given by (Kane, 1963; Morgan, 1965)

$$\sigma_E = \frac{e^2}{4\pi\varepsilon} \left[2\pi \left(N_D + N_A \right) r_s \right]^{1/2} \tag{16.28}$$

where r_s is the screening radius given by (see Eq. 16.7b)

$$r_s = \left[-\left(\frac{dn}{dV} \right)_{V=0} \frac{e}{\varepsilon} \right]^{-1/2} . \tag{16.29}$$

Note that the square-root dependence of the energy fluctuation, σ_E, on the doping concentration, $N_D + N_A$, is a result of the Poisson distribution assumed for impurities. The Poisson distribution can be replaced by a gaussian distribution, if the energy fluctuations are small. Small fluctuations occur if the random variation of the number of donor atoms within a spherical volume defined by the screening radius is much smaller than the average number of donors within the sphere (Kane, 1963), *i. e.*

$$\sqrt{(N_D + N_A)\, r_s^3} \quad << \quad (N_D + N_A)\, r_s^3 . \tag{16.30}$$

The fluctuation of the potential, *e. g.* the conduction band edge potential, can then be expressed in terms of a gaussian distribution. The probability for the conduction band edge energy to occur at an energy E_C, is given by

$$p(E_C) = \frac{1}{\sqrt{2\pi}\ \sigma_E} \, e^{-\frac{1}{2}\left(\frac{E_C}{\sigma_E} \right)^2} \tag{16.31}$$

where we defined the mean energy of the conduction band edge as $E_C = 0$ and σ_E is given by Eq. (16.28). The unperturbed density of states in the conduction band is given by

$$\rho_{DOS}(E) = \frac{1}{2\pi^2} \left(\frac{2m^*}{\hbar^2} \right)^{3/2} \sqrt{E - E_C} . \tag{16.32}$$

The density of states in a perturbed potential is obtained by a summation over all locations in space, *i. e.* by an integral over the probability distribution of the band edge

$$\rho_{DOS,Kane} = \frac{1}{2\pi^2} \left(\frac{2m^*}{\hbar^2} \right)^{3/2} \int_{-\infty}^{E} \sqrt{E - E_C} \, \frac{1}{\sqrt{2\pi}\ \sigma_E} \, e^{-\frac{1}{2}\left(\frac{E_C}{\sigma_E} \right)^2} \, dE_C \tag{16.33}$$

where $\rho_{DOS,Kane}$ is the density of states according to Kane (1963), or briefly the Kane function. The density of states given by Eq. (16.33) is valid for the conduction band. The corresponding density of states of the valence band can be obtained by replacing E_C by $-E_V$, E by $-E$, and the electron effective mass by the heavy-hole effective mass.

The parabolic density of states, the gaussian probability distribution of the band edge, and the Kane function are shown in ***Fig***. 16.10. The density of states shown in ***Fig***. 16.10(c) tails into the gap. Tail states have energies lower than the average band-edge energy. It is important to note that the density of states of the Kane function does not change the *bandgap* energy, *i. e.* the spatially *averaged* position of the conduction band edge does not change. The tailing of the band into the forbidden gap changes the density of state drastically in the vicinity of the band edge. However, the density of states is practically unchanged for energies $E - E_C >> \sigma_E$, as can be easily inferred from Eq. (16.33). The Kane function is of great practical use for the simple,

quantitative description of band edges. For example, Casey and Stern (1976) used the Kane function to calculate absorption and spontaneous emission in doped GaAs.

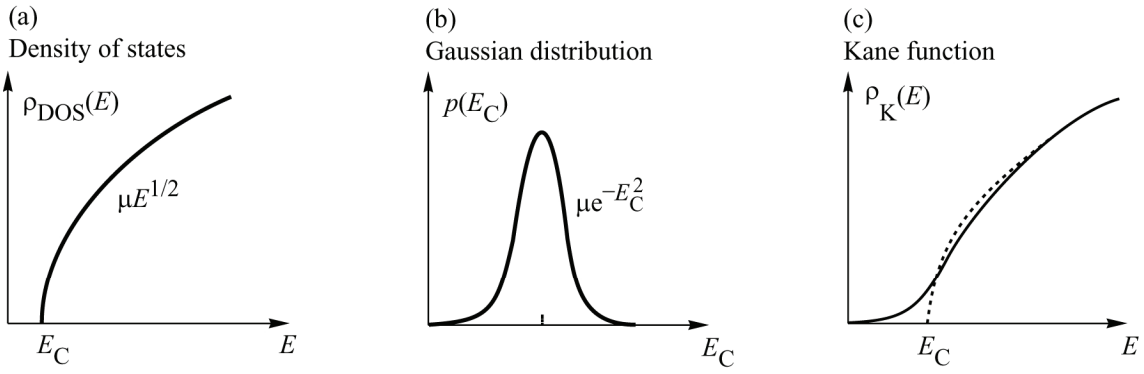


Fig. 16.10. (a) Density of states of a parabolic, spherical valley. (b) Gaussian probability distribution. (c) The Kane function describes the density of states of a parabolic band including tail states at $E < E_C$ which arise from potential fluctuations caused by randomly distributed impurities.

The density of states according to Kane (Eq. 16.33) overestimates the extent of band tailing since tunneling of carriers through the potential barriers is not taken into account. In addition, the quantization of carriers in the potential minima is not taken into account in the Kane model. However, in the limit of carriers with a large effective mass, the Kane function applies, since quantum effects can be neglected for such heavy carriers.

Halperin and Lax (1966, 1967) and Lax and Halperin (1966) calculated the density of states in band tails taking into account (*i*) the quantization of carriers in the potential minima as well as (*ii*) tunneling of carriers through potential barriers. The importance of carrier quantization can be estimated by comparing the magnitude of the potential fluctuations with the quantization energy of carriers. The former is given by Eq. (16.28) for poissonian (random) impurity distributions. The latter can be estimated from

$$E_Q \approx \frac{\hbar^2}{2m^* r_s^2} \tag{16.34}$$

where r_s is the screening radius. The density of states calculated by Halperin and Lax (1966) is shown in ***Fig***. 16.11 along with the density of states of the unperturbed bands and the density of states of the Kane model for GaAs with doping concentrations of $N_A = 1.1 \times 10^{19}$ cm^{-3} and $N_D = 9 \times 10^{18}$ cm^{-3} (Casey and Stern, 1976). The density of states calculated according to the Kane model shown in ***Fig***. 16.12 was used to interpolate between the unperturbed band and the Halperin–Lax result. The parameter σ_E (see Eq. 16.33) was adjusted in order to obtain a good fit between the two functions.

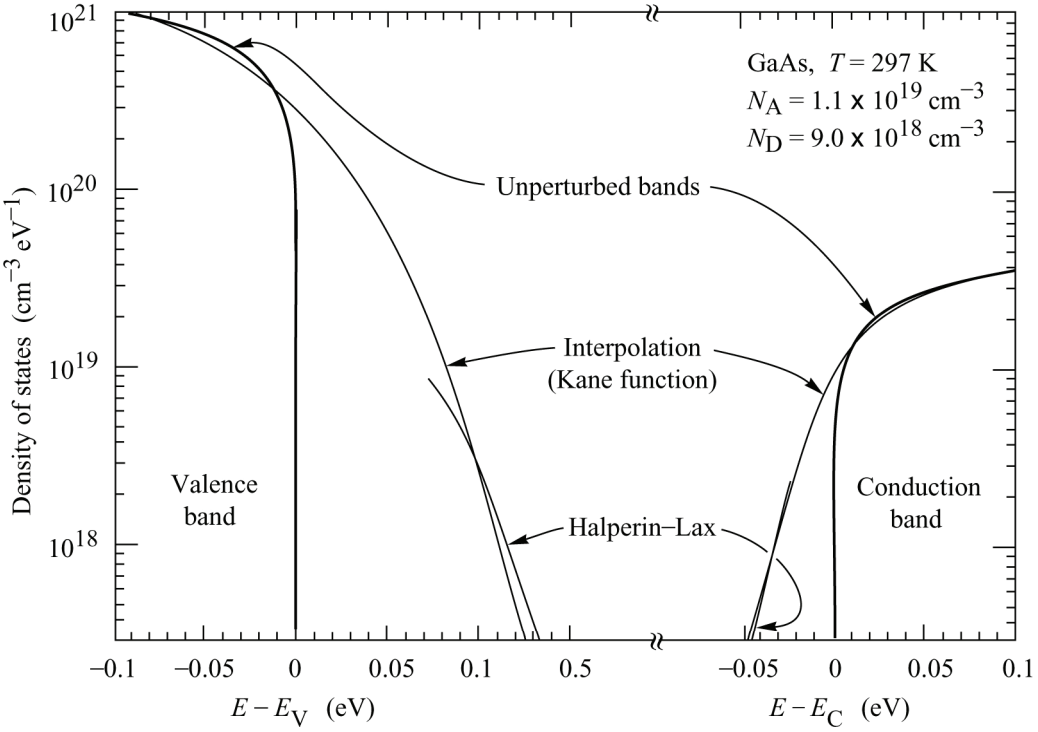


Fig. 16.11. Densities of states in the conduction and valence band versus energy for GaAs with a net acceptor concentration of 2×10^{18} cm^{-3}. The curves show the densities of states in the unperturbed bands and the densities of states in the band tails calculated by Halperin and Lax. The curves which join them are interpolated Kane functions (after Casey and Stern, 1976).

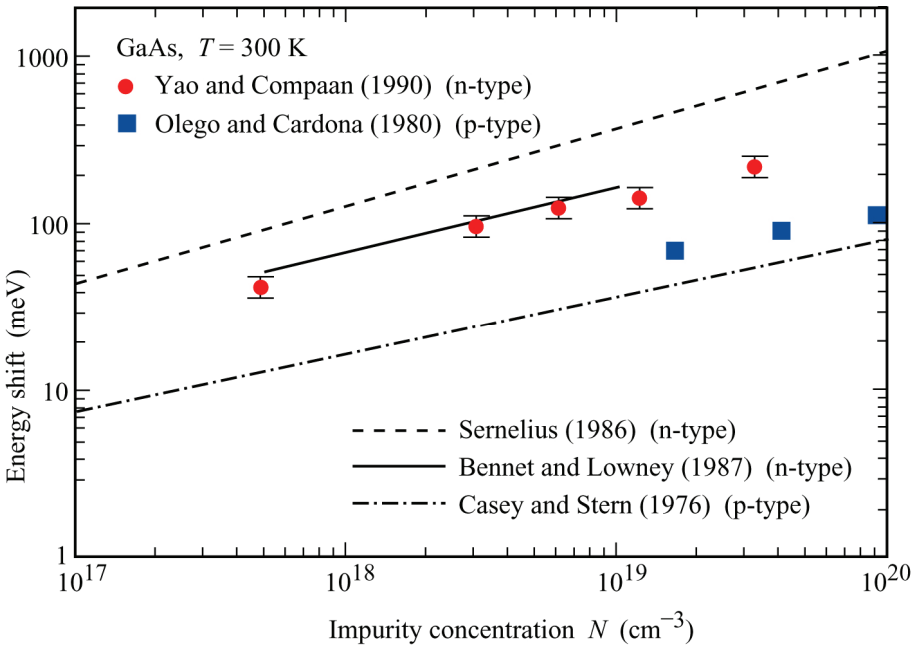


Fig. 16.12. Bandgap narrowing for n-type and p-type GaAs as a function of impurity concentration (after Yao and Compaan, 1990).

16.6 Bandgap narrowing

At high doping concentrations the bandgap energy of semiconductors decreases. The magnitude of the reduction increases with doping concentration and is usually referred to as ***bandgap narrowing, bandgap shrinkage***, or ***bandgap renormalization***. There are many reasons for the reduction of the band gap which have been reviewed by Abram *et al.* (1978). The most important reasons for bandgap narrowing are many-body effects of free carriers which lower the electron energies as compared to a non-interacting carrier system. Many-body effects describe the interaction of free carriers. The interaction becomes important at small carrier-to-carrier distances, *i. e.* at high free carrier concentrations. Electrons can interact with each other either by their long-range Coulomb potential or via their spin. We first consider coulombic electron–electron interactions. Consider an electron which is added to a highly doped, neutral semiconductor. When the electron is added to the semiconductor, other electrons in the vicinity of the added electron spatially redistribute in order to reduce the long-range coulombic interaction energy. The energy of the electron added to the semiconductor is *reduced* by the redistribution of neighboring electrons. Many-body effects are thus related to screening, *i. e.* the spatial redistribution of carriers in the presence of potential perturbations. Carriers can also interact via their spin (Mahan, 1990). Due to the Fermion nature of electrons, each volume element in phase space can be occupied by at most two electrons with opposite spin (Pauli principle). Electrons with like spin have a repulsive interaction while electrons with opposite spin have an attractive interaction. If electrons were distributed uniformly throughout the crystal, the attractive and repulsive energies would exactly cancel each other. However, due to the interaction, electrons with like spin tend to stay away from each other and electrons with opposite spin tend to stay closer. As a result, the interaction energy reduces the total energy of the electron system.

Many-body interactions can occur between free carriers and between free carriers and ionized impurities. Such interactions are called carrier–carrier and carrier–impurity interactions. Thus, bandgap narrowing occurs in highly doped semiconductors as well as in undoped, but highly excited semiconductors with high free carrier concentrations. We summarize the many-body interactions in n- and p-type semiconductors as follows:

(*i*) *Electron–electron interactions*. The repulsive and attractive interactions of electrons with like and opposite spin and the long-range coulombic interactions lead to a net attractive term. As a consequence, the conduction band edge is lowered (*i. e.* lowering of gap energy).

(*ii*) *Electron–donor interactions*. The interaction of electrons and ionized donors is attractive and leads to another lowering of the conduction band edge.

(*iii*) *Hole–hole interactions*. The spin interaction energy and the coulombic interaction result in a net attractive energy and an increase in the valence band edge (*i. e.* lowering of gap energy).

(*iv*) *Hole–acceptor interactions*. The interaction between holes and donors is attractive and leads to an increase of the valence band edge (*i. e.* decrease in gap energy).

(*v*) *Electron-hole interactions*. Highly excited semiconductors have a large concentration of electrons and holes. Interaction effects lead to a reduction of the energy gap in highly excited semiconductors.

Next we calculate the magnitude of bandgap renormalization due to electron–electron interaction effects and follow the calculation of Haug and Schmitt-Rink (1985). At room temperature, the magnitude of bandgap renormalization can be approximated by the classical self energy of an electron interacting with an electron gas, *i. e.* interacting with its own polarization field. Using the classical analogy, we define the self energy of an electron interacting with an electron gas as

$$\Delta E_g \approx e \lim_{r \to 0} \left[V_s(r) - V(r) \right] \tag{16.35}$$

where $V(r)$ is the Coulomb potential and $V_s(r)$ is the screened Coulomb potential of an electron in an electron gas. Equation (16.35) thus represents the change in electrostatic energy of an electron before and after the electron gas has spatially redistributed itself to reduce the Coulomb interaction energy. In other words, ΔE_g describes the energy that the electrons gain by avoiding each other. Assuming that the screened potential is of the Yukawa form, we obtain

$$\Delta E_g \approx -\frac{e^2}{4\pi\varepsilon r_s} \tag{16.36}$$

The screening radius, r_s, is given by the Debye and the Thomas–Fermi radii in non-degenerate and degenerate semiconductors, respectively. Insertion of the screening radii into Eq. (16.36) yields the bandgap renormalization energies

$$\Delta E_g = -\frac{e^3 \sqrt{n}}{4\pi\varepsilon^{3/2}\sqrt{kT}} \qquad \text{(Debye)} \tag{16.37a}$$

$$\Delta E_g = -\frac{e^3 \sqrt{m^* (3n)^{1/3}}}{4\pi^{5/3}\varepsilon^{3/2}\hbar} \qquad \text{(Thomas–Fermi)} \tag{16.37b}$$

The momentum-dependence of the renormalization energy is relatively small (Wolf, 1962; Haug and Schmitt-Rink, 1985). If the momentum-dependence is neglected, bandgap renormalization produces a *rigid downward movement* of the conduction band dispersion (or, in p-type semiconductors, a rigid upward movement of the valence band). Finally, bandgap renormalization also occurs in undoped but highly excited semiconductors. Brinkman and Rice (1973) examined the effect of electron–hole interactions in highly excited semiconductors and showed that bandgap renormalization occurs due to electron–hole interactions. Phenomenological expressions for the carrier–carrier interaction and the carrier–impurity interaction were given by Mahan (1980) and Landsberg *et al.* (1985). Assuming that the interaction energies are proportional to the carrier–carrier distance, the change in energy gap follows a 1/3 power of the doping concentration, that is

$$\Delta E_g \propto (N_D)^{1/3} \tag{16.38}$$

where full activation of the donors was assumed, *i. e.* $n = N_D$. Similar considerations are valid for p-type semiconductors. In this case carrier-carrier and carrier–acceptor interactions must be considered. Note that the phenomenological dependence of ΔE_g on n given in Eq. (16.38), *i. e.* $\Delta E_g \propto n^{1/3}$, is weaker than the Debye result, $\Delta E_g \propto n^{1/2}$ (see Eq. 16.37a), but stronger than the Thomas–Fermi result, $\Delta E_g \propto n^{1/6}$ (see Eq. 16.37b).

Experimental and theoretical results of bandgap narrowing as a function of doping concentration are shown in ***Fig***. 16.12 for n-type and p-type doping of GaAs (Yao and Compaan, 1990). The graph includes experimental data for n-type (Yao and Compaan, 1990), and p-type GaAs (Olego and Cardona, 1980), as well as theoretical data for n-type (Sernelius, 1986; Bennet and Lowney, 1987) and p-type GaAs (Casey and Stern, 1976). The data of ***Fig***. 16.12 shows that bandgap narrowing can assume quite large values, such as 200 meV, at high n-type doping concentrations. Furthermore, the change in bandgap energy follows the $N^{1/3}$ dependence predicted by Eq. (16.38). The magnitudes of bandgap narrowing can be expressed by the following simple formulas:

$$\text{n-type GaAs:} \qquad \Delta E_g \text{ (meV)} \approx -6.6 \times 10^{-5} \sqrt[3]{N_D \text{ (cm}^{-3})} \tag{16.39}$$

$$\text{p-type GaAs:} \qquad \Delta E_g \text{ (meV)} \approx -2.4 \times 10^{-5} \sqrt[3]{N_A \text{ (cm}^{-3})} \ . \tag{16.40}$$

These equations represent a fit to the experimental data shown in ***Fig***. 16.12. As expected, the bandgap narrowing for n-type GaAs is larger as compared to p-type GaAs (Yao and Compaan, 1990). Bandgap narrowing was studied to a smaller extent in other III–V semiconductors. An approximate formula for bandgap narrowing in n-type InP was given by Böer (1990)

$$\text{n-type InP:} \qquad \Delta E_g \text{ (meV)} \approx -2.25 \times 10^{-5} \sqrt[3]{N_D \text{ (cm}^{-3})} \tag{16.41}$$

where it is assumed that all donors are active ($n = N_D$). Further phenomenological expressions and parameters for other III–V semiconductors were given by Jain *et al.* (1990).

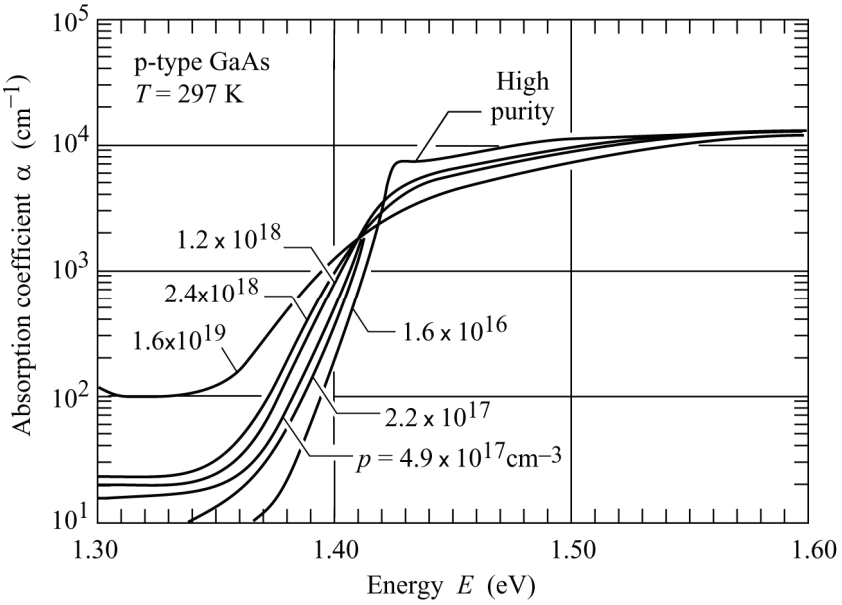


Fig. 16.13. Absorption coefficient for p-type GaAs at 297 K for different doping concentrations (after Casey and Stern, 1976).

The Burstein–Moss shift and bandgap narrowing are two phenomena which cause the Fermi level of highly doped semiconductors to change in opposite directions. While bandgap narrowing causes the Fermi level of an n-type semiconductor to decrease, the Burstein–Moss shift is due to an increase of the Fermi level. Absorption measurements allow one to measure

$E_F - E_V$ and $E_C - E_F$ for n-type and p-type semiconductors, respectively. In n-type GaAs, the Burstein–Moss shift prevails, resulting in a blue-shift of the absorption edge (see Sect. on *Burstein–Moss shift*). In p-type GaAs, bandgap narrowing prevails and results in a red-shift of the absorption edge. The absorption coefficient of p-type GaAs is shown in ***Fig***. 16.13 for different p-type doping levels (Casey and Stern, 1976). High-purity GaAs exhibits a rapidly decreasing absorption coefficient at the fundamental gap. However, as the doping concentration increases, the absorption below the gap energy increases as well. Furthermore, the absorption coefficient decreases less rapidly as compared to the high-purity GaAs. Both characteristics indicate that bandgap narrowing and the formation of tail states dominate the near-band-edge optical absorption rather than the Burstein–Moss shift. The strongly different absorption characteristics of p-type (see ***Fig***. 16.13) and n-type GaAs (see ***Fig***. 16.6) are due to the heavier hole mass as compared to the effective electron mass. The Burstein–Moss shift is inversely proportional to the carrier effective mass, which results in less band filling in p-type GaAs.

References

Abram R. A., Rees, G. J., and Wilson, B. L. H. "Heavily Doped Semiconductors and Devices *Advances in Physics* **27**, 799 (1978)

Adler D., in *Handbook on Semiconductors* **1**, edited by T. S. Moss and W. Paul (North-Holland, Amsterdam, 1982)

Ando T., Fowler A. B., and Stern F. "Electronic properties of two-dimensional systems" *Reviews of Modern Phys*ics **54**, 437 (1982)

Austin I. G. and Mott N. F. "Polarons in crystalline and non-crystalline materials" *Advances in Physics*, **18**, 41 (1969)

Bennet, H. S and Lowney, J. R. "Models for heavy doping effects in gallium arsenide" *Journal of Applied Physics* **62**, 521 (1987)

Böer K., *Survey of Semiconductor Physics* (van Nostrand Reinhold, New York, 1990)

Brinkman W. F., and Rice T. M. "Electron-Hole Liquids in Semiconductors" *Physical Review B***7**, 1508 (1973)

Burstein, E. "Anomalous Optical Absorption Limit in InSb" *Physical Review* **93**, 632 (1954)

Casey H. C., Sell D. D., and Wecht K. W. "Concentration dependence of the absorption coefficient for *n*- and *p*-type GaAs between 1.3 and 1.6 eV" *Journal of Applied Physics* **46**, 250 (1975)

Casey Jr. H. C. and Stern F. "Concentration-dependent absorption and spontaneous emission of heavily doped GaAs" *Journal of Applied Physics* **47**, 631 (1976)

Dapkus P. D., Holonyak Jr. N., Rossi J. A., Williams L. V. and High D. A. "Laser transition and wavelength limits of GaAs" *Journal of Applied Physics* **40**, 3300 (1969)

Debye O. P. and Hückel E. "The theory of electrolytes" *Physikalische Zeitschrift* **24**, 185, 305 (1923)

Debye P. and Conwell E. M. "Electrical properties of N-type germanium" *Physical Review* **93,** 693 (1954)

Deppe D. G., Gerrard N. D. Pinzone C. J. Dupuis R. D. and Schubert E. F. "Quarter-wave Bragg reflector stack of InP-In/sub 0.3/Ga/sub 0.47/As for 1.65 μm wavelength" *Applied Physics Letters* **56**, 315 (1990)

Efros A. L. and Shklovskii B. I. "Coulomb gap and low temperature conductivity of disordered systems" *Journal of Physics* C**8**, L49 (1975)

Flügge S., *Practical Quantum Mechanics* **1 (**Springer Verlag, Berlin, 1971)

Halperin B. I. and Lax M. "Impurity-Band Tails in the High-Density Limit: I. Minimum Counting Methods" *Physical Review* **148**, 722 (1966)

Halperin, B. I. and Lax, M. "Impurity-Band Tails in the High-Density Limit: II. Higher Order Corrections" *Physical Review* **153,** 802 (1967)

Haug H. and Schmitt-Rink S. "Electron theory of the optical properties of laser-excited semiconductors" *Progress in Quantum Electronics*, **9**, 3 (1984)

Haug H. and Schmitt-Rink S. "Basic mechanisms of the optical nonlinearities of semiconductors near the

band edge**"** *Journal of the Optical Society of America B***2**, 1135 (1985)
Hubbard J. "Electron correlations in narrow energy bands," *Proceedings of the Royal Society (London) Series A***276**, 238 (1963)
Jain S. C. McGregor J. M. and Roulston D. J. "Band-gap narrowing in novel III-V semiconductors" *Journal of Applied Physics* **68**, 3747 (1990)
Kane E. O. "Thomas-Fermi Approach to Impure Semiconductor Band Structure" *Physical Review* **131**, 79 (1963)
Landsberg P. T., Neutroschel A., Lindholm F. A., and Sah C. T. "A model for band-gap shrinkage in semiconductors with application to silicon" *Physica Status Solidi*, *B***130**, 255 (1985)
Lax M. and Halperin B. I. "Title unknown to EFS" *Journal of the Physical Society of Japan* **21** (Supplement), 218 (1966)
Mahan G. D. "Energy gap in Si and Ge: Impurity dependence" *Journal of Applied Physics* **51**, 2634 (1980)
Mahan G. D., *Many-Particle Physics* (Plenum, New York, 1990)
Morgan T. N. "Broadening of Impurity Bands in Heavily Doped Semiconductors" *Physical Review* **139**, A343 (1965)
Moss T. S., *Optical Properties of Semiconductors* (Academic Press, New York, 1961)
Mott N. F., *Conduction in Non-Crystalline Materials* (Clarendon, Oxford, 1987)
Mott N. F., *Metal-Insulator Transitions* (Taylor & Francis, London, 1990)
Olego D. and Cardona A. "Photoluminescence in heavily doped GaAs: I. Temperature and hole-concentration dependence" *Physical Review B* **22**, 886 (1980)
Pollak M. and Knotek M. L. "Correlation effects in hopping conduction: A treatment in terms of multielectron transitions" *Physical Review B***9**, 664 (1974)
Sernelius B. E. "Band-gap shifts in heavily doped *n*-type GaAs" *Physical Review B***33**, 8582 (1986)
Shockley W. "Problems related to p-n junctions in Silicon" *Solid-State Electronics* **2**, 35 (1961)
Shklovskii B. I. and Efros A. L. *Electronic Properties of Doped Semiconductors* (Springer Verlag Verlag, Berlin, 1984)
Stern F. "Polarizability of a Two-Dimensional Electron Gas", *Physical Review Letters* **18**, 546 (1967)
Stillman G. E., Cook L. W., Roth T. J., Low T. S. and Skromme B. J., in *GaInAsP Alloy Semiconductors*, edited by T. P. Pearsall (John Wiley and Sons, New York, 1982)
Verie C. "Title unknown to EFS" *Proc. Int. Conf. on II-VI Compounds* p. 1124 (Benjamin, New York, 1967)
Wolf P. A. "Theory of the Band Structure of Very Degenerate Semiconductors" *Physical Review* **126**, 405 (1962)
Yao H. and Compaan A. "Plasmons, photoluminescence, and band-gap narrowing in very heavily doped n-GaAs" *Applied Physics Letters* **57**, 147 (1990)

17

Band diagrams of heterostructures

17.1 Band diagram lineups

In a semiconductor heterostructure, two different semiconductors are brought into physical contact. In practice, different semiconductors are "brought into contact" by epitaxially growing one semiconductor on top of another semiconductor. To date, the fabrication of heterostructures by epitaxial growth is the cleanest and most reproducible method available. The properties of such heterostructures are of critical importance for many heterostructure devices including field-effect transistors, bipolar transistors, light-emitting diodes and lasers.

Before discussing the lineups of conduction and valence bands at semiconductor interfaces in detail, we classify heterostructures according to the alignment of the bands of the two semiconductors. Three different alignments of the conduction and valence bands and of the forbidden gap are shown in ***Fig***. 17.1. ***Figure*** 17.1(a) shows the most common alignment which will be referred to as the ***straddled alignment*** or "Type I" alignment. The most widely studied heterostructure, that is the GaAs / $Al_xGa_{1-x}As$ heterostructure, exhibits this straddled band alignment (see, for example, Casey and Panish, 1978; Sharma and Purohit, 1974; Milnes and Feucht, 1972). ***Figure*** 17.1(b) shows the ***staggered lineup***. In this alignment, the steps in the valence and conduction band go in the same direction. The staggered band alignment occurs for a wide composition range in the $Ga_xIn_{1-x}As$ / $GaAs_ySb_{1-y}$ material system (Chang and Esaki, 1980). The most extreme band alignment is the ***broken gap alignment*** shown in ***Fig***. 17.1(c). This alignment occurs in the InAs / GaSb material system (Sakaki *et al*., 1977). Both the staggered lineup and the broken-gap alignment are called "Type II" energy band alignments.

At the semiconductor interface of the heterostructure, the energies of the conduction and valence band edges change. The magnitudes of the changes in the band-edge energies are critically important for many semiconductor devices.

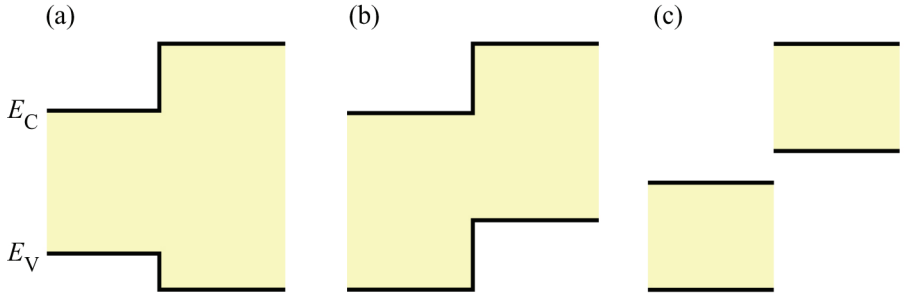


Fig. 17.1. Types of energy band lineups: (a) straddled or "Type I" lineup, (b) staggered or "Type II" lineup, and (c) broken or "Type III" lineup.

There have been numerous attempts and models to predict and calculate the energy ***band offsets*** in semiconductor heterostructures (Anderson, 1962; Harrison, 1977, 1980, 1985; Frensley and Kroemer, 1977; Kroemer, 1985; Ruan and Ching, 1987; Van de Walle, 1989; Van de Walle and Martin, 1986; Tersoff 1984, 1985, 1986; Harrison and Tersoff, 1986). The different models have been reviewed by Kroemer (1985) and by Ruan and Ching (1987). The authors showed that the agreement between the theoretical and experimental band offsets varies for the different approaches. However, none of the theoretical approaches can reliably predict the band offsets of

all semiconductor heterostructure combinations. Here, we restrict ourselves to a few empirical rules and fundamental theoretical concepts which will be useful for the understanding of heterojunction band discontinuities.

Linear superposition of atomic-like potentials

We first discuss the model of the linear superposition of atomic-like potentials developed by Kroemer (1975, 1985). He pointed out that the problem of theoretically understanding the relative alignment of bands is the problem of determining the relative alignment of the two periodic potentials of the two semiconductors forming the heterostructure. Once the periodic potential of a semiconductor or of a heterostructure is known, the energy bands can be calculated.

The periodic potential of a semiconductor can be viewed as a linear superposition of the overlapping atomic-like potentials as shown in ***Fig***. 17.2. Near the atomic nuclei, the atomic-like potentials resemble the potentials inside the free atoms. However, a reconfiguration of the valence electrons occurs when initially isolated atoms form a lattice of atoms. The atomic potentials in a solid state atomic lattice will be different from the atomic potentials of isolated atoms. Therefore, the potentials in a solid-state lattice are designated as atomic-*like* potentials.

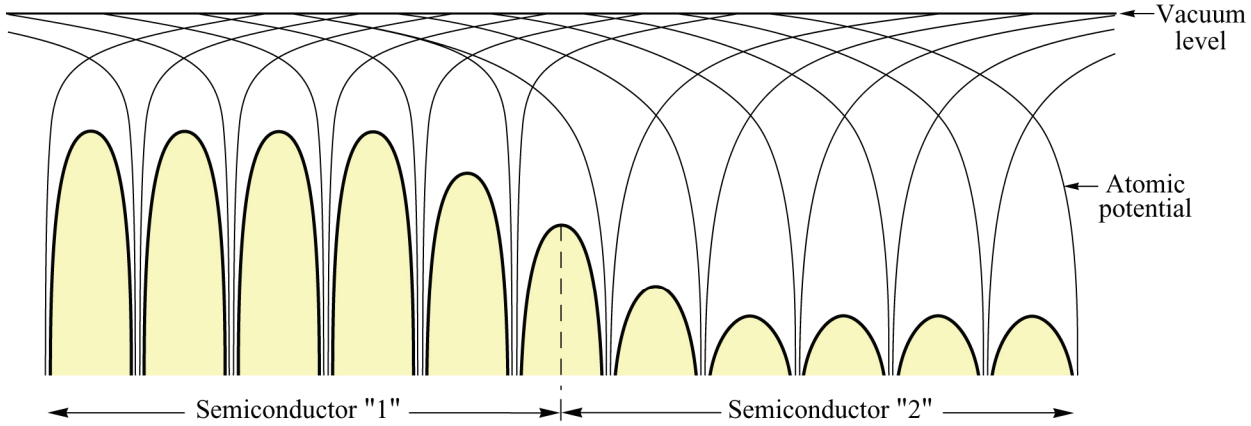


Fig. 17.2. Atomic potentials in the vicinity of two semiconductors "1" and "2". Within each semiconductor, all atomic potentials are identical. The resulting crystal potential is obtained by the superposition of all atomic potentials.

In the simplest atomic theory of band lineups, the unmodified atomic-like potentials would be superimposed throughout the entire structure. In the intimate vicinity of the interface, the potential would contain contributions from atoms from both sides of the interface, as shown in ***Fig***. 17.2. However, deep inside either of the two semiconductors, the periodic potential would be unaffected by the atomic-like potentials of the other semiconductor. In such a model, the lineup of the periodic potentials is well defined. The band lineups are then also well defined, and the only problems are those of the computational technique used to calculate the bandstructure from the periodic potential. Although the model of the superposition of atomic-like potentials is very instructive, the ability of this model to predict offsets between semiconductors is very limited (Kroemer, 1985).

We next consider the transition region between the two semiconductors, namely the abruptness of this transition. Atomic and atomic-like potentials are *short-range* potentials. They decay exponentially and have completely vanished after only a few inter-atomic distances, as schematically shown in ***Fig***. 17.2. As a result of the short-range nature of the atomic potential,

the transition region in which the potential has intermediate values will be very thin, *i. e*. at most just a few atomic layers thick. Assuming that the bands closely follow the periodic potential, the transition of bands from the bulk structure in one semiconductor to the bulk structure in the other semiconductor will also occur within a very thin layer. The model of the linear superposition of atomic-like potentials therefore demonstrates that the transition region for chemically abrupt interfaces is very thin, namely just a few atomic layers thick. The free carrier de Broglie wavelength is much longer than the transition region. Therefore, *the potential and band transition region at the interface between two semiconductor can be considered to be abrupt* for chemically abrupt semiconductor interfaces. In other words, the electronic transition between two semiconductors is (nearly) as abrupt as the chemical transition.

Van de Walle and Martin (1986) calculated the atomic potentials of Si and Ge in Si / Ge heterostructures. The calculation indeed confirmed that the transition region from the Ge bulk periodic potential to the Si bulk periodic potential is very thin, namely just two monolayers thick. Assuming that the energy bands closely follow the periodic potential, the transition region from the Ge bulk band diagram to the Si bulk band diagram is also just a few atomic monolayers thick. Hence, the periodic potential and energy band calculations of Van de Walle and Martin clearly confirmed the assumption of Kroemer that the transition region in chemically abrupt semiconductor heterostructures is just a few monolayers thick.

The electron affinity model

The electron affinity model is the oldest model invoked to calculate the band offsets in semiconductor heterostructures (Anderson, 1962). This model has proven to give accurate predictions for the band offsets in several semiconductor heterostructures, whereas the model fails for others. We first outline the basic idea of the electron affinity model and then discuss the limitations of this model.

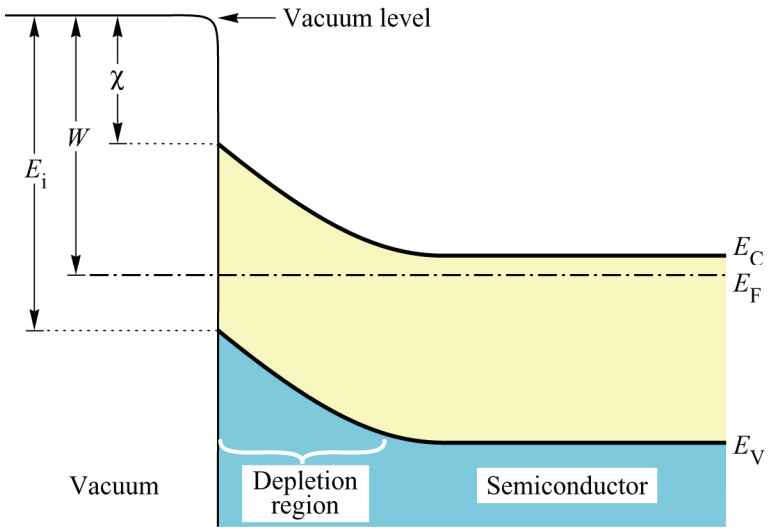


Fig. 17.3. Electron affinity χ, work function W, and ionization energy E_i of a semiconductor. The electron affinity is measured from the bottom of the conduction band at the semiconductor surface, the work function from the Fermi level, and the ionization energy from the top of the valence band at the surface.

The band diagram of a semiconductor-vacuum interface is shown in ***Fig***. 17.3. Near the surface, the n-type semiconductor is depleted of free electrons due to the pinning of the Fermi level near the middle of the forbidden gap at the semiconductor surface. Such a pinning of the Fermi level at the surface occurs for most semiconductors. The energy required to move an electron from the semiconductor to the vacuum surrounding the semiconductor depends on the initial energy of the electron in the semiconductor. Promoting an electron from the bottom of the

conduction band to the vacuum beyond the reach of image forces requires work called the ***electron affinity*** χ. Lifting an electron from the Fermi level requires work called the ***work function*** W, which is defined the same way in semiconductors as it is in metals. Finally, raising an electron from the top of the valence band requires the ***ionization energy*** E_i. This energy is measured by photoionization experiments, in which semiconductors are illuminated by monochromatic light with a variable wavelength. The longest wavelength at which photoionization occurs defines the ionization energy.

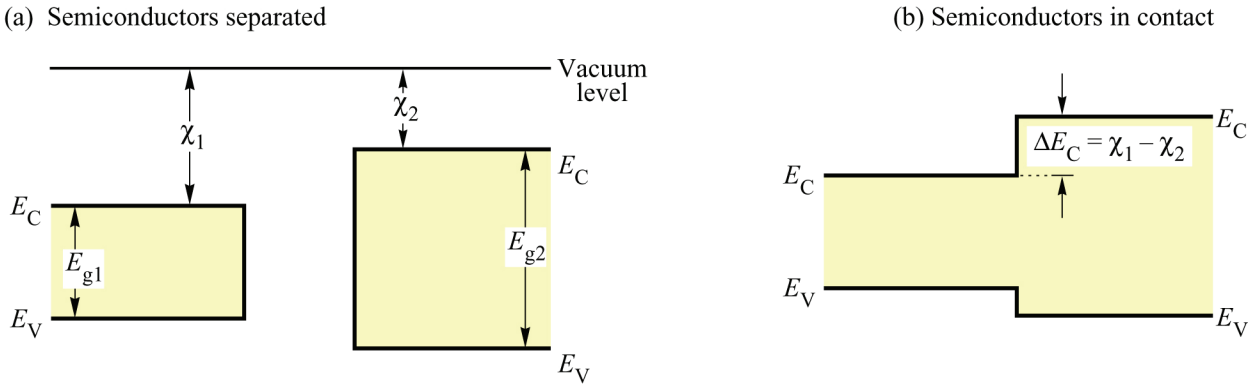


Fig. 17.4. Band diagrams of (a) two separated semiconductors and (b) two semiconductors in contact. The semiconductors have a band gap energy of E_{g1} and E_{g2} and an electron affinity of χ_1 and χ_2.

Next consider that two semiconductors are brought into physical contact. The two semiconductors are assumed to have an electron affinity of χ_1 and χ_2 and a bandgap energy of E_{g1} and E_{g2}, respectively, as illustrated in ***Fig***. 17.4. Near-surface band bending and the effect of image forces have been neglected in the figure. The electron affinity model is based on the fact that the energy balance of an electron moved from the vacuum level to semiconductor "1", from there to semiconductor "2", and from there again to the vacuum level must be zero, that is $\chi_1 - \Delta E_c - \chi_2 = 0$ or

$$\boxed{\Delta E_c = \chi_1 - \chi_2} \tag{17.1}$$

The valence band discontinuity then follows automatically as

$$\Delta E_v = E_{g2} - E_{g1} - \Delta E_c \tag{17.2}$$

Note that Eqs. (17.1) and (17.2) are valid only if the potential steps caused by atomic dipoles at the semiconductor surfaces and the heterostructure interfaces can be neglected. In this case, the knowledge of the electron affinities of two semiconductors provides the band offsets between these two semiconductors. Shay *et al*. (1976) concluded that the influence of dipole layers at semiconductor surfaces change the values of the electron affinity by about only 1%. Therefore, the authors argued, the electron affinity rule is indeed applicable to semiconductor heterostructures.

The electron affinity model has successfully explained the band discontinuities of several semiconductor heterostructures. In the InAs / GaSb material system, the electron affinity rule correctly predicts a broken-gap alignment (Gobeli and Allen, 1966; Kroemer, 1985). The highly

asymmetric lineup of InAs / GaAs heterostructures is also predicted well (Kroemer, 1985). In the Si / Ge heterostructure system, the electron affinity model predicts $\Delta E_c = 0.12$ eV and $\Delta E_v = 0.33$ eV in reasonable agreement with experimental data (Kroemer, 1985). Shay *et al*. (1976) and Phillips (1981) used the electron affinity rule to calculate ΔE_c in CdS / InP heterostructures and found excellent agreement with their experimental data.

Despite the reasonable agreement between theory and experiment, the electron affinity model suffers from several conceptual problems which have been pointed out by Kroemer (1985). *First*, surface dipole layers affect the measurement of the electron affinity. Generally, all semiconductor surface undergo surface reconstruction, *i. e*. a rearrangement of atoms on the semiconductor surface in order to reduce the total energy of the semiconductor surface. Such a surface reconstruction includes frequently the outward or inward displacement of surface atoms. As a result, electrostatic dipole layers are formed which will change the measured electron affinity. At semiconductor-semiconductor interfaces, the interface reconstruction will be clearly different than the surface reconstruction. As a consequence, the magnitude of interface dipoles will be different. Therefore, the measurement of χ is influenced by surface effects and the measured values of χ will not be meaningful for semiconductor heterostructures, unless the influence of surface and interface dipoles is negligible small, or if the surface dipoles are identical to the interface dipoles. Both possibilities are unlikely. However, Shay *et al*. (1976) pointed out that the influence of surface dipoles is very small for most semiconductor surfaces. *Second*, electron correlation effects also influence the measured values of the electron affinity (Kroemer, 1985). When one electron is taken from a semiconductor and promoted to the vacuum level, the remaining electrons will rearrange themselves in order to reduce the total energy of the electron system. Such correlation effects are due to coulombic repulsion between electrons but also due to quantum-mechanical exchange effects (essentially the Pauli exclusion principle). Generally, the magnitude of correlation effects is small. Due to the dipole and correlation effects, the applicability of the electron affinity rule is limited to semiconductors in which these effects are negligibly small.

It is useful to recall that the electron affinity model was invoked by Schottky (1938, 1940) explain the barrier heights of metal-semiconductor contacts also called Schottky contacts. Schottky proposed that the barrier height be given by the difference in the work function in the metal and the electron affinity of the semiconductor, *i. e*. $W - \chi$. However, it is well known, that the Schottky model clearly fails to explain the barrier heights in metal-semiconductor contacts. Subsequently, Bardeen (1947) showed, the important role of interface states whose energy is within the forbidden gap. Bardeen showed that interface dipoles caused by charged interface states determine the barrier height of metal-semiconductor contacts and that the difference $W - \chi$ does not play a significant role. In lattice-matched semiconductor-semiconductor junctions, the influence of interface dipoles cannot be possibly as large as it is in metal-semiconductor junctions. Lattice-matched semiconductor heterostructures have highly ordered atomic transitions between the two semiconductors with relatively little atomic and electronic reconstruction. Therefore, the electron affinity model is expected to provide much better results for semiconductor-semiconductor junctions than it does for metal-semiconductor junctions. This expectation is indeed confirmed by experimental results.

Common anion rule

Many compound semiconductor heterostructures consist of two compounds which share a common anion element. For example in AlGaAs / GaAs heterostructures, As is the anion element on both sides of the heterostructure. It is a well established fact that the valence band wave functions evolve mainly from the atomic wave function of anions and the conduction band

wave functions evolve mainly from the atomic wave functions of cations (see, for example, Harrison, 1980). Hence, the valence band structure of different semiconductors with the same anion element will be similar. Furthermore, *the valence band offsets of compound semiconductors with the same anion element is generally smaller than the conduction band offset*. This rule is clearly confirmed in the material system $Al_xGa_{1-x}As$ / GaAs where $\Delta E_c / \Delta E_g \approx 2/3$ and $\Delta E_V / \Delta E_g \approx 1/3$ for direct-gap range of $Al_xGa_{1-x}As$ ($x \leq 0.45$). The common anion rule also works well for GaAs / InAs heterostructures in which $\Delta E_c / \Delta E_v \approx 5/1$ (Kowalczyk *et al*., 1982).

Harrison atomic orbital model

Harrison (1977, 1980, 1985) developed a theory based on atomic orbitals to predict band offsets in semiconductor heterostructures. Kroemer (1985) compared the Harrison atomic orbital model and other models with experiments and he arrived at the conclusion that the Harrison model gives very good overall agreement with experimental band offsets.

The basis of the Harrison model is the *linear combination of atomic orbitals* of a very small group of atoms which is then used to calculate the band structure. The band structure calculation would be correct if the true atomic-like potentials and energy eigenfunctions of the atoms forming the semiconductor would be known. Because the atomic-like potentials and eigen energies of the atoms in the crystal lattice are unknown, Harrison simply takes as unperturbed atomic energy values the theoretical values of *free* atoms. Hence, the Harrison model is clearly an approximation. In this model, several more approximations are employed for the calculation of the matrix elements coupling the relevant atomic states between nearest neighbors. For further discussion, the reader is referred to the literature (Harrison, 1977, 1980, 1985, Kroemer, 1985).

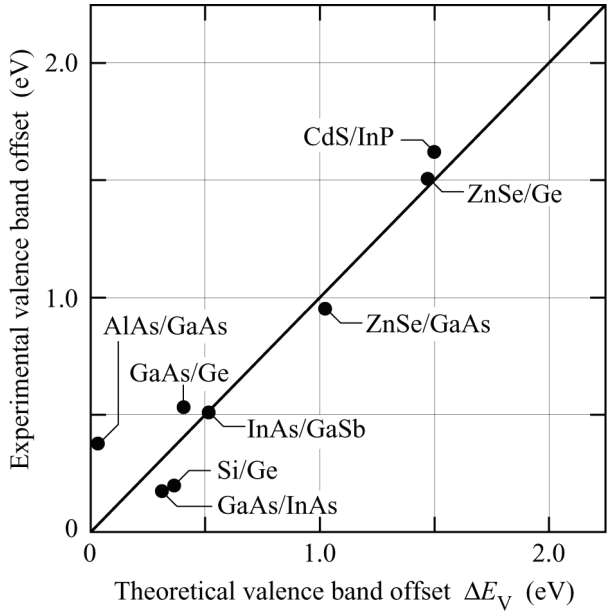


Fig. 17.5. Comparison of experimental valence band offsets with theoretical valence band offsets calculated by the Harrison atomic orbital theory. The AlAs/GaAs value is extrapolated from $Al_{0.30}Ga_{0.70}As$/GaAs assuming $\Delta E_V/\Delta E_g = 1/3$.

A comparison between Harrison's theoretical and experimental valence band offsets is shown in ***Fig***. 17.5. The data used in the figure was compiled by Kroemer (1985) except the value for the GaAs / $Al_xGa_{1-x}As$ where $\Delta E_v = 0.32\ \Delta E_g$ has been used, consistent with more recent results (Pfeiffer *et al*., 1991). ***Figure*** 17.5 displays a very good overall agreement between experiment and the Harrison atomic orbital model.

The effective dipole model

As we have already stated above, any dipole charges at the heterointerface will change the heterostructure band discontinuity. These dipole charges are due to the locally different atomic and electronic structure at the heterointerface as compared to the bulk atomic structure of either semiconductor. As a result of the different atomic environment at the heterointerface, valence electrons of atoms at the interface will move from their bulk equilibrium positions to new equilibrium positions. Hence, atomic dipoles are formed due to the new charge distribution at the heterointerface.

Ruan and Ching (1987) calculated heterostructure band offsets based on (i) the electron affinity model and (ii) by taking into account atomic dipoles at the interface which cause an additional shift of the band discontinuity. The authors pointed out that interface dipoles are neglected in Anderson's electron affinity model. If no net charge is transferred between the two semiconductors forming the heterojunction, then the Anderson model gives the correct band offset. (We do not consider here the difficulties in obtaining the correct electron affinities χ_e, but simply assume that they are known. Ruan and Ching used "average values of those experimental data which are judged to be current and reliable".)

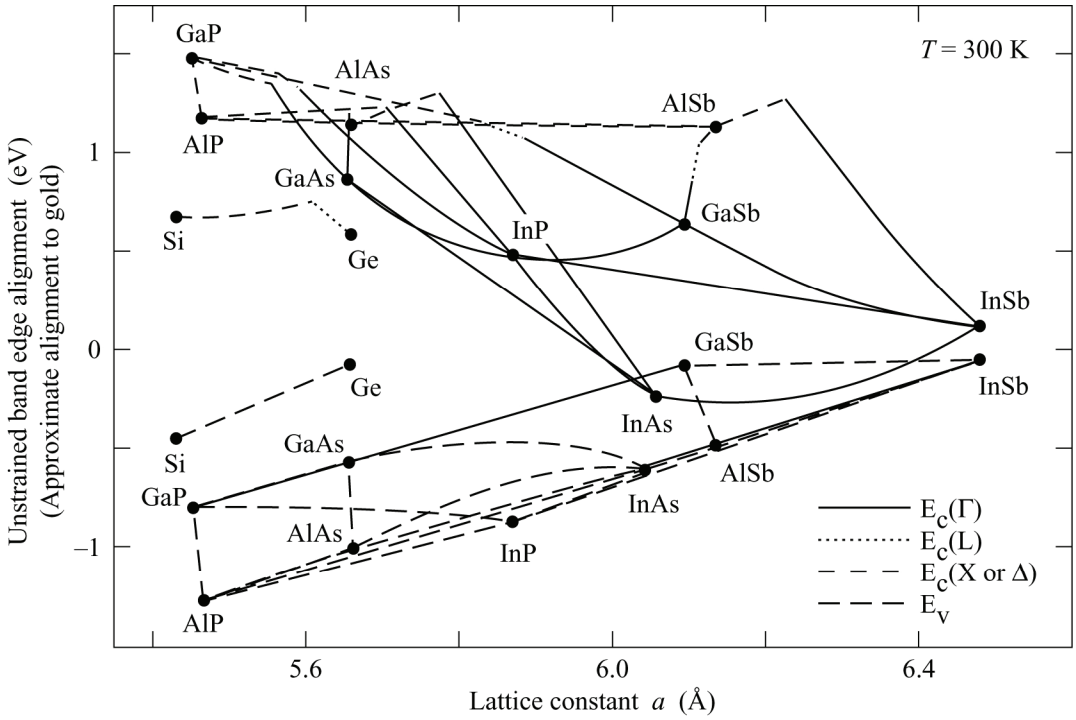


Fig. 17.6. Band edges as a function of the lattice constant. The zero energy point represents the approximate Fermi level (located mostly in the forbidden gap) of the gold / semiconductor Schottky contact (after Tiwari and Frank, 1992).

To calculate the charge transfer between the two semiconductors forming the heterojunction, Ruan and Ching (1987) assume that the two valence bands are misaligned, that is the valence band edges of the two semiconductors have different energies. Electrons with an effective mass m^* in the valence band of one semiconductor will tunnel into the forbidden regions of the other semiconductor. The dipole charge is calculated by integrating over the exponentially decaying charge distribution of electrons tunneling into into the forbidden gaps of the adjoining semiconductor. Using this method to calculate the band offsets between semiconductors, Ruan and Ching (1987) calculated nearly all conceivable heterostructure band offsets. A comparison

revealed that the theoretical band offsets of Ruan and Ching differs, on average, only about 0.1 eV from the available experimental data.

Experimental data of band offsets between different semiconductors are given in Table 3. The table includes data for elemental as well as binary and ternary compound semiconductors. Tiwari and Frank (1992) used experimental data of band offsets in order to plot the band edges of semiconductors as a function of the lattice constant. The plot, shown in ***Fig***. 17.6 relies on the experimental observation that the band offset from material "A" to material "B" plus the offset from material "B" to material "C" is equal to the band offset from material "A" to material "C". This *linearity* of offsets is consistent with the electron affinity model, and this property allows one to predict band alignments of any semiconductor heterostructure.

Material system A / B	E_g^A (eV)	E_g^B (eV)	ΔE_V (eV)	$\Delta E_V / \Delta E_g$ (absolute value)	Remarks
Si / Ge	1.12	0.67	–0.16 to –0.40	0.35 to 0.89	(a)
Si / GaP	1.12	2.25	+0.80	0.71	(b)
Si / GaAs	1.12	1.42	+0.05	0.17	(b)
Si / GaSb	1.12	0.72	–0.05	0.12	(b)
Si / ZnSe	1.12	2.70	+1.25	0.79	(b)
Si / CdTe	1.12	1.52	+0.75	1.87	(b)
Ge / AlAs	0.67	2.15	+0.92	0.62	(b)
Ge / GaAs	0.67	1.42	+0.25 to +0.65	0.33 to 0.87	(b)
Ge / InP	0.67	1.34	+0.64	0.95	(b)
AlAs / GaAs	2.15	1.42	–0.40	0.55	(c)
$Al_{0.3}Ga_{0.7}As$ / GaAs	1.79	1.42	–0.12	0.32	(d)
AlSb / GaSb	1.61	0.72	–0.4	0.45	(b)
GaAs / InAs	1.42	0.36	–0.17	0.16	(b)
GaAs / ZnSe	1.42	2.70	+0.96 to +1.10	0.75 to 0.86	(b)
GaSb / InAs	0.72	0.36	+0.46	1.28	(b)
InP / CdS	1.34	2.42	+1.63	1.51	(b)
$Al_{0.48}In_{0.52}As$ / $Ga_{0.47}In_{0.53}As$	1.45	0.75	–0.21	0.30	(e)
$Ga_{0.52}In_{0.48}P$ / GaAs	1.88	1.42	–0.23	0.50	(f)
$Al_{0.48}In_{0.52}As$ / InP	1.45	1.34	+1.19	1.73	(g)
$Ga_{0.47}In_{0.53}As$ / InP	0.75	1.34	+0.40	0.68	(g)

(a) after Ruan and Ching (1987). Van de Walle and Martin (1986) showed that ΔE_V depends strongly on strain.
(b) after Ruan and Ching (1987)
(c) indirect gap AlAs, after Ruan and Ching (1987)
(d) direct gap $Al_xGa_{1-x}As$, after Pfeiffer *et al*. (1991) and after Menendez *et al*. (1986)
(e) after Peng *et al*. (1986) and after Sugiyama *et al*. (1986)
(f) Rao *et al*. (1987)
(g) after Tiwari and Frank (1992)

Table 17.1: Bandgap energies and valence band offsets of semiconductor heterostructures "A / B". The valence band offset ΔE_V is positive, if the top of the valence band of semiconductor "A" is higher than that of semiconductor "B".

17.2 Boundary conditions at heterointerface

So far we have seen how the energy bands in semiconductors evolve and how these bands align in semiconductor heterostructures. We have also seen that the transition from one band diagram to the band diagram of another semiconductor is very abrupt in a chemically abrupt semiconductor heterostructure. In this section we will discuss the transition of other physical quantities at the heterointerface. These transitions follow a number of rules and these rules are called the *boundary conditions* of the heterointerface. Below, the boundary conditions will be summarized. A heterointerface is schematically illustrated in ***Fig***. 17.7. The interface is located in the plane $z = 0$ of a cartesian coordinate system. The semiconductors "1" and "2" have a dielectric permittivity, magnetic permeability, and effective mass of ε_1, μ_1, m_1^*, and ε_2, μ_2, m_2^*, respectively.

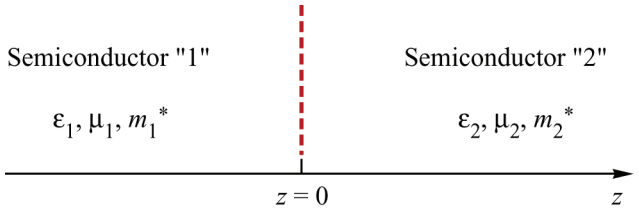


Fig. 17.7. Boundary between two semiconductors "1" and "2" which have a permittivity of ε_1, and ε_2, and an effective mass m_1^* and m_2^*, respectively.

The first boundary condition considered here concerns the Fermi level. *The **Fermi level** is constant across a heterointerface under thermal equilibrium conditions*. The Fermi level is defined as the energy at which electronic states are populated with a probability of one half. We next assume a heterointerface, in which the Fermi level on one side of the heterointerface is, at a given time, different from the Fermi level on the other side of the heterointerface. As a consequence, electrons will transfer from the semiconductor with the higher Fermi level to the semiconductor with the lower Fermi level, where they can occupy states at lower energy. Thus, the Fermi level rises in the semiconductor with the initially lower Fermi level. (Also, the Fermi level decreases in the semiconductor with the initially higher Fermi level.) The transfer of electrons continues, until the Fermi level is the same on both sides of the heterointerface. Thus, under thermal equilibrium conditions, the Fermi level is constant across heterointerfaces.

Four electrodynamic boundary conditions must be satisfied at heterointerfaces. The general boundary conditions for electric and magnetic fields were derived in Chap. 2. For completeness, these boundary conditions are summarized as follows: The ***magnetic boundary condition*** states that the tangential component of the magnetic field, $\mathcal{H}_t$, and the normal component of the magnetic induction, $\mathcal{B}_n$, are constant across interfaces. The ***electric boundary condition*** states that the tangential component of the electric field, $\mathcal{E}_t$, and the normal component of the dielectric displacement, $\mathcal{D}_n$, are constant across interfaces.

The latter boundary condition, $\mathcal{D}_n$ = const, is now used to derive another "boundary condition", namely the ***charge neutrality condition***. We denote the normal component of the dielectric displacement at the interface as D_{1n} and D_{2n} in semiconductor "1" and "2", respectively. Furthermore we assume that the dielectric displacement vanishes for sufficiently large distances from the interface. Then, using Gauss's equation, the boundary condition $\mathcal{D}_{1n} = \mathcal{D}_{2n}$ can be written as

$$\mathcal{D}_{1n} = \int_{z=-\infty}^{z=0} \rho(z)\,dz = -\int_{z=0}^{z=\infty} \rho(z)\,dz = \mathcal{D}_{2n}\,. \qquad (17.3)$$

The equation states that the net charge on one side of the heterointerface $z \le 0$, left-hand side of Eq. (17.3) must be equal to the negative net charge on the other side of the heterointerface

($z \geq 0$, right-hand side of Eq. (17.3)). All charges must be taken into account in Eq. (17.3) including free-carrier accumulation layer charges, free-carrier inversion layer charges, and depletion layer charges. Charges at heterointerfaces are caused by the transfer of carriers from one semiconductor across the heterointerface to the other semiconductor. All of these charges remain in close vicinity of the heterointerface. Thus the condition of charge neutrality of Eq. (17.3) can be stated as: *There is not net charge in the vicinity of heterointerfaces*.

We finally discuss the boundary conditions for the quantum-mechanical wave function at heterointerfaces. The interface is located in the plane $z = 0$ and we are only interested in the z dependence of the wave function $\psi(z)$. In the chapter entitled "Resume of quantum mechanical principles", the boundary conditions for $\psi(z)$ is given by

$$\boxed{\psi_1(z \to -0) \quad = \quad \psi_2(z \to +0)} \tag{17.4}$$

That is, the wave function is continuous at the interface.

The boundary condition for the derivative of the wave function is given by

$$\boxed{\frac{1}{m_1^*}\frac{\mathrm{d}\psi_1}{\mathrm{d}z}\bigg|_{z\to -0} \quad = \quad \frac{1}{m_2^*}\frac{\mathrm{d}\psi_2}{\mathrm{d}z}\bigg|_{z\to +0}} \tag{17.5}$$

The proof of this equation is given in the chapter entitled "Resume of quantum mechanical principles".

17.3 Graded gap structures

In regular semiconductor heterostructures, the chemical transition from one semiconductor to another semiconductor structure is abrupt. In the preceding section, we have seen that the periodic potential and the band diagram are nearly as abrupt as the chemical transition. That is, the transition of the periodic potential and of the band diagram occur within a few atomic layers of a chemically abrupt semiconductor heterostructure. In graded heterostructures, the chemical transition from one semiconductor to another semiconductor is intentionally graded. In this section, the properties of such graded heterostructures are discussed.

Assume two semiconductors "A" and "B" that are chemically miscible. The mixed compound, also called ***semiconductor alloy***, is designated by the chemical formula A_xB_{1-x}, where x is the ***mole fraction*** of semiconductor "A" in the mixed compound. The mole fraction x is also designated as the chemical ***composition*** of the compound A_xB_{1-x}. Most semiconductors of practical relevance are completely miscible. Assume further that the gap energy of semiconductor "A" and "B" are different and that the bandgap energy depends on the composition. The dependence of the forbidden-gap energy on the composition x is usually expressed in terms of a parabolic (linear plus quadratic) dependence. The gap energy of the alloy A_xB_{1-x} is then given by

$$E_g^{AB} \quad = \quad x\,E_g^{A} \;+\; (1-x)\,E_g^{B} \;+\; x\,(1-x)\,E_b \tag{17.6}$$

where the first two summands describe the linear dependence of the gap and the summand $x\,(1-x)\,E_b$ describes the quadratic dependence of the gap. The parameter E_b is called the ***bowing parameter***. For some semiconductor alloys, *e. g.* $(AlAs)_x(GaAs)_{1-x}$, the bowing parameter is vanishingly small. The bandgap of the alloy is then given by

$$E_g^{AB} = x\,E_g^{A} + (1-x)\,E_g^{B} \tag{17.7}$$

Equations (17.6) and (17.7) are valid for homogeneous bulk semiconductors. However, the validity of the equations is not limited to bulk semiconductors. They also apply to the local bandgap of graded structures. We have seen in the preceding section that the atomic potentials and the energy bands closely follow the composition in a chemically abrupt heterojunction. Accordingly, the band edges and the gap energy will follow the chemical composition of graded semiconductors.

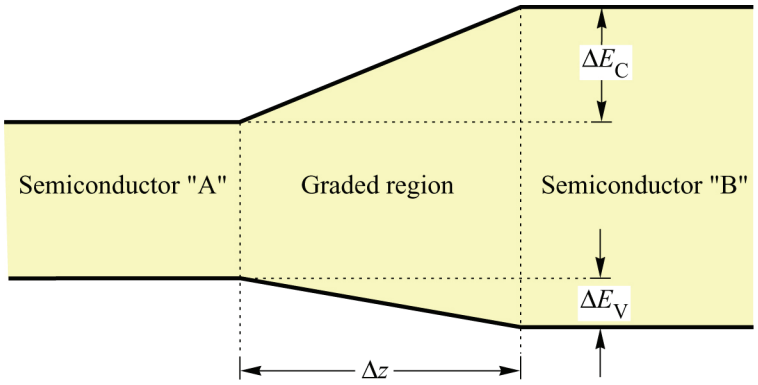


Fig. 17.8. Linearly graded region between a semiconductor "A" and "B" creating a quasi-electric field in the graded transition region.

The band diagram of a linearly graded semiconductor heterostructure is illustrated in ***Fig***. 17.8. The figure shows a narrow-gap semiconductor "A", a wide-gap semiconductor "B", and a linearly graded transition region "A_xB_{1-x}" with thickness Δz. It is assumed that ΔE_c, ΔE_V, and ΔE_g depend linearly on the composition x. Graded gap semiconductor structures were first considered by Kroemer (1957). He showed that the changes of the band edge energies with position can be understood as ***quasi-electric fields***. The quasi-electric field of the band diagram shown in ***Fig***. 17.8 is, in the conduction band, given by

$$|\mathcal{E}_C| = \Delta E_C / (e\,\Delta z) \tag{17.8}$$

In the valence band it is given by

$$|\mathcal{E}_V| = \Delta E_V / (e\,\Delta z) \tag{17.9}$$

Figure 17.8 reveals that the electric fields in the conduction band and in the valence band have opposite polarization. Therefore, electrons and holes are driven in the *same* direction (to the left-hand side of figure). This cannot be achieved by real electric fields in which electrons and holes are *always* driven in opposite directions. Due to this difference, Kroemer (1957) designated the fields occurring in graded semiconductor structures as ***quasi-electric fields***. Kroemer envisioned several different cases of graded gap structures and pointed out that in some graded gap structures, electrons and holes are pulled in the same direction, as discussed for the band diagram shown in ***Fig***. 17.8. In other graded gap structures, one of the bands could, *e. g*. the valence band, may be flat, while the other band could have a quasi-electric field. Kroemer also envisioned graded-gap heterobipolar transistors which enhance the minority carrier transport through the base.

Kroemer's conjecture that the quasi-electric fields exert forces on free carriers was experimentally verified by Levine *et al*. (1982, 1983). The authors showed that the quasi-electric fields act on one carrier type just like regular electric fields of the same magnitude would. That

is, the drift velocity relation $v = \mu \mathcal{E}$ (where $\mathcal{E}$ is the quasi-electric field) was confirmed by a modified Shockley-Haynes experiment (Levine *et al*., 1982).

There are many very interesting applications for graded gap structures including graded-base heterobipolar transistors (Kroemer, 1957; Miller *et al*. 1983; Hayes *et al*. 1983), the elimination of heterojunction band discontinuities (Schubert *et al*. 1992), and several graded gap structures for optoelectronic applications such as photodetectors. Graded-gap detectors have been reviewed by Capasso (1984, 1986).

Let us consider some experimental results of alloy semiconductors. The energy gap of the unstrained Si_xGe_{1-x} was analyzed as a function of composition by Weber and Alonso (1989) using low-temperature photoluminescence. Analytical expressions were given for the lowest energy gap as a function of the composition. For $x \leq 0.85$, the X band is the lowest conduction band minimum. The energy gap of unstrained Si_xGe_{1-x} is given by

$$E_g(x) = \left(1.155 - 0.43\,x + 0.206\,x^2\right)\text{eV} \qquad \text{for} \qquad x \leq 0.85 \qquad (17.10a)$$

For $x > 0.85$, the L band is the lowest conduction band minimum. The energy gap is then given by

$$E_g(x) = \left(2.010 - 1.270\,x\right)\text{eV} \qquad \text{for} \qquad x > 0.85 \qquad (17.10b)$$

For strained Si_xGe_{1-x} grown on Si substrates, Lang *et al*. (1985) showed that the degenerate valence band splits into two bands. The energy gap between the lowest conduction band and the highest valence band is then given by

$$E_g(x) = \left(1.155 - 0.65\,x + 0.22\,x^2\right)\text{eV} \qquad \text{for} \qquad x \leq 0.70 \qquad (17.10c)$$

This equation is an analytical expression of the low-temperature (90 K) photoluminescence data of Lang *et al*.

The energy gaps of ternary III–V alloy semiconductors have been compiled by Swaminathan and Macrander (1991). The data is summarized in Table 4. The energy gaps of quaternary III–V and for II-VI semiconductors will not be summarized here. The interested reader is referred to the literature (Pearsall, 1982; Landolt-Börnstein, 1987).

In graded semiconductor structures, the composition of the semiconductor is varied. This variation in chemical composition is not only accompanied by a change of the bandgap energy, but also by a change in the lattice constant. The change in lattice constant is, for all semiconductor alloys, governed by Vegard's law. Consider a semiconductor "A" with a lattice constant a_0^A and a semiconductor "B" with the lattice constant a_0^B. Then the lattice constant of the alloy A_xB_{1-x} is given by ***Vegard's law*** which states

$$\boxed{a_0^{AB} = a_0^A x + a_0^B (1-x)} \qquad (17.11)$$

For most graded semiconductor structures, it is imperative that the lattice constant does not change as the composition of the alloy is varied. Such structures are called ***lattice-matched*** graded semiconductors. If semiconductors are not lattice matched, graded semiconductors. If semiconductors are not lattice matched, microscopic defects occur when the composition is varied. These defects degrade the quality, *e. g*. the radiative efficiency, of the semiconductor.

The relationship between the gap energy, the corresponding wavelength, and the lattice constant of group-IV and group III–V semiconductors is shown in ***Fig***. 17.9 (Tien, 1985). A similar plot is shown in ***Fig***. 17.10 for II-VI semiconductors (Feldman *et al*., 1992). The two

plots allow one to select lattice-matched semiconductors with the desired bandgap energy. ***Figure*** 17.9 reveals that the lattice constant of the material system $Al_xGa_{1-x}As$ / GaAs does not change, as the composition x is varied. This advantageous properly allows one to easily grow lattice-matched graded structures with this material system.

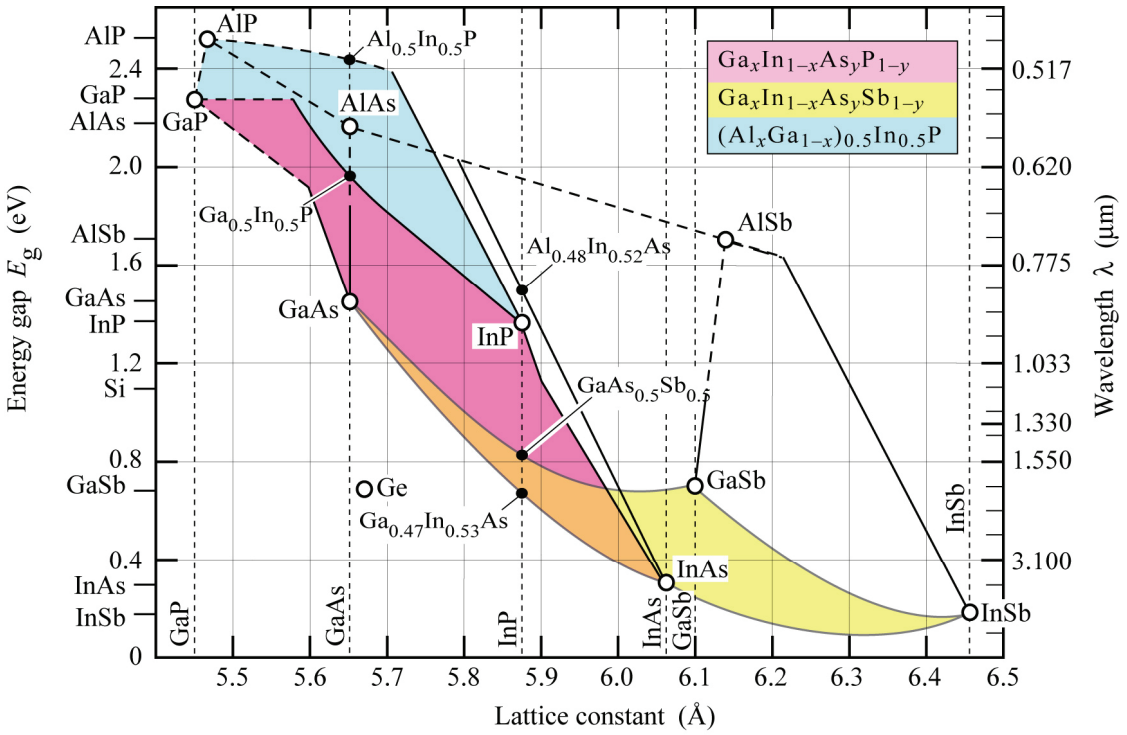


Fig. 17.9. Lattice constant versus energy gap at room temperature for various III–V semiconductors and their alloys (after Tien, 1985).

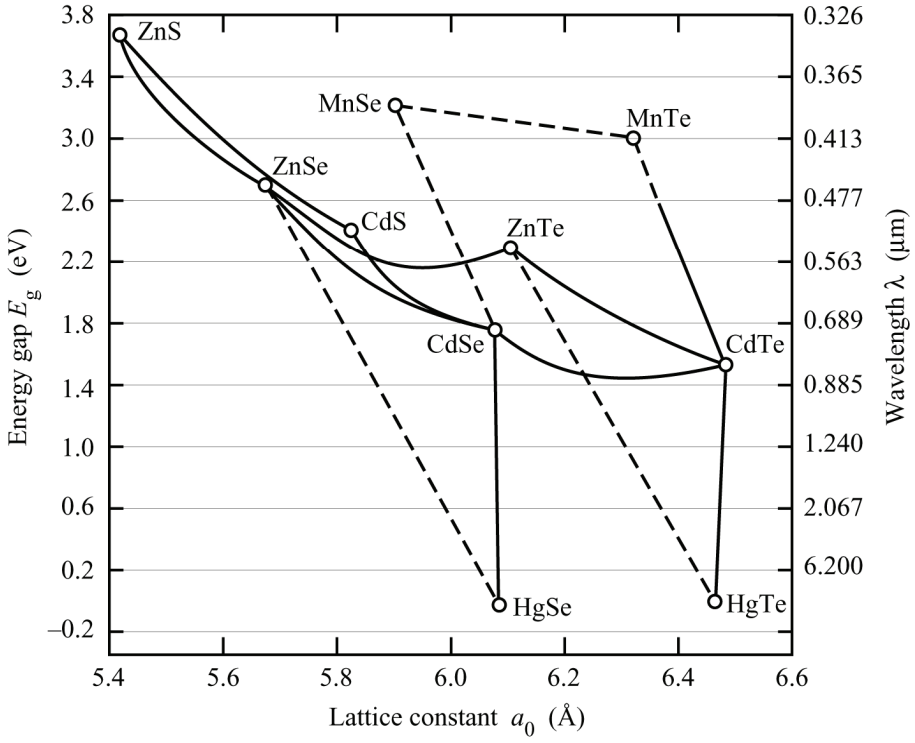


Fig. 17.10. Lattice constant versus energy gap at room tempera-ture for various II–VI semiconductors and their alloys (after Feldman *et al.*, 1992).

Alloy	Direct Energy Gap E_Γ (eV)	Indirect Energy Gap	
		E_X (eV)	E_L (eV)
$Al_xIn_{1-x}P$	$1.34 + 2.23x$	$2.24 + 0.18x$	
$Al_xGa_{1-x}As$	$1.424 + 1.247x$ ($x < 0.45$) $1.424 + 1.087x + 0.438x^2$	$1.905 + 0.10x + 0.16x^2$	$1.705 + 0.695x$
$Al_xIn_{1-x}As$	$0.36 + 2.35x + 0.24x^2$	$1.8 + 0.4x$	
$Al_xGa_{1-x}Sb$	$0.73 + 1.10x + 0.47x^2$	$1.05 + 0.56x$	
$Al_xIn_{1-x}Sb$	$0.172 + 1.621x + 0.43x^2$		
$Ga_xIn_{1-x}P$	$1.34 + 0.511x + 0.604x^2$ ($0.49 < x < 0.55$)		
$Ga_xIn_{1-x}As$	$0.356 + 0.7x + 0.4x^2$		
$Ga_xIn_{1-x}Sb$	$0.172 + 0.165\,x + 0.413x^2$		
GaP_xAs_{1-x}	$1.424 + 1.172x + 0.186x^2$		
$GaAs_xSb_{1-x}$	$0.73 - 0.5x + 1.2x^2$		
InP_xAs_{1-x}	$0.356 + 0.675x + 0.32x^2$		
$InAs_xSb_{1-x}$	$0.18 - 0.41x + 0.58x^2$		

Table 17.2: Compositional dependence of the gap energy in ternary III–V semiconductors at room temperature.

17.4 Semiconductor heterostructures

- Lattice matching required for low defect density
- This is particularly important for minority carrier devices
- This is not so important for majority carrier devices
- Ideal: Heterostructures are formed by semiconductors with the same crystal structure and the same lattice constant: An example is $Al_xGa_{1-x}As$ on GaAs
- Often: Mismatched structures result in misfit dislocations defects which act as recombination centers. An example is GaN on sapphire
- Diagrams of energy gap-versus-lattice-constant for of different semiconductors

References

Anderson R. K. "Experiments on Ge-GaAs heterojunctions" *Solid State Electronics* **5**, 341 (1962)

Bardeen, J. "Surface States and Rectification at a Metal Semi-Conductor Contact" *Physical Review* **71**, 717 (1947)

Capasso, F. "New multilayer and graded gap optoelectronic and high speed devices by band gap engineering" *Surface Science,* **142,** 513 (1984)

Capasso, F. "Compositionally Graded Semiconductors and their Device Applications" *Ann. Rev. Mater. Sci* **16**, 263 (1986)

Casey H. C. and Panish M. B. *Heterostructure Lasers, Part B*, p. 16 (Academic Press, New York, 1978)

Chang L. L. and Esaki L. "Electronic properties of InAs-GaSb superlattices [and MBE]" *Surface Science* **98,** 70 (1980)

Feldman R. D., Lee D., Partovi A., Stanley R. P., Johnson A. M., Zucker J. E., Glass A. M., and Hegarty J. "Growth, optical, and optoelectronic properties of CdZnTe/ZnTe multiple quantum wells" *Critical Reviews in Solid State Material Science* **17,** 477 (1992z)

Frensley W. R and Kroemer H. "Theory of the energy-band lineup at an abrupt semiconductor heterojunction" *Physical Review* B**16**, 2642 (1977)

Gobeli G. W. and Allen F. G., in *Semiconductors and Semimetals* edited by R. K. Willardson and A. C. Beer (Academic Press, New York, 1966)

Harrison W. A. "Elementary theory of Heterojunctions" *Journal of Vacuum Science and Technology* **14**, 1016 (1977)

Harrison W. A., *Electronic Structure and the Properties of Solids: The Physics of the Chemical Bond* (Freeman, San Francisco, 1980)

Harrison W. A. "Theory of band line-ups" *Journal of Vacuum Science and Technology B***3**, 1231 (1985)

Harrison W. A. and Tersoff J. "Tight-binding theory of heterojunction band lineups and interface dipoles" *Journal of Vacuum Science and Technology B***4** 1068 (1986)

Hayes J. R., Capasso F., Gossard A. C., Malik R. J., and Wiegmann W. "Bipolar transistor with graded band-gap base" *Electronics Letters* **19**, 410 (1983)

Kowalczyk S. P., Schaffer W. J., Kraut E. A., and Grant R. W. "Determination of the InAs/GaAs(100) heterojunction band discontinuities by x-ray photoelectron spectroscopy (XPS)" *Journal of Vacuum Science and Technology* **20**, 705 (1982)

Kroemer H. "Quasi-Electric and Quasi-Magnetic Fields in Non-Uniform Semiconductors," *RCA Review*, **18**, 332 (1957)

Kroemer, C. "Problems in the theory of heterojunction discontinuities**"** *Critical Reviews in Solid State Sciences*, **5**, 555 (1975)

Kroemer H. in *Molecular Beam Epitaxy and Heterostructures* edited by L. L. Chang and K. Ploog (Martinus Nijhoff, Dordrecht, 1985).

Landolt-Börnstein, *Numerical Data and Functional Relationships in Science and Technology, New Series* edited by O. Madelung, Vol. **22,** Sub-volume A (Springer Verlag, Berlin, 1987)

Lang D. V., People R., Bean J. C. and Sergent A. M. "Measurement of the band gap of $Ge_xSi_{1-x}Si$ strained-layer heterostructures" *Applied Physics Letters,* **47,**1333 (1985)

Levine B. F., Tsang W. T., Bethea C. G. and Capasso F. "Electron drift velocity measurement in

compositionally graded $Al_xGa_{1-x}As$ by time-resolved optical picosecond reflectivity" *Applied Physics Letters* **41**, 470 (1982)

Levine B. F., Bethea C. G. Tsang W. T., Capasso F., and Thornbar K. K. "Measurement of high electron drift velocity in a submicron, heavily doped graded gap $Al_xGa_{1-x}As$ layer" *Applied Physics Letters* **42**, 769 (1983)

Miller D. L., Asbeck P. M., Anderson R. J. and Eisen F. H. "AlGaAs/GaAs heterojunction bipolar transistors with graded composition in the base" *Electronics Letters* **19**, 367 (1983)

Milnes A. G. and Feucht D. L. *Heterojunctions and Metal-Semiconductor Junctions* (Academic Press, New York, 1972)

Pearsall T. P., editor *GaInAsP Alloy Semiconductors* (John Wiley and Sons, New York, 1982)

Phillips J. C. "Partial dislocations, columnar growth, clustering, and pinhole formation in ultra-thin film semiconductor heterostructures" *Journal of Vacuum Science and Technology* **19**, 545 (1981)

Ruan Y.-C. and Ching W. Y. "An effective dipole theory for band lineups in semiconductor heterojunctions" *Journal of Applied Physics* **62**, 2885 (1987)

Sakaki J., Chang L. L. Ludeke R., Chang C.-A., Sai-Halasz G. A., and Esaki L. "$In_{1-x}Ga_xAs$/ $GaSb_{1-y}As_y$ heterojunctions by molecular-beam epitaxy" *Applied Phys Letters*, **31**, 211 (1977)

Schottky W. "Semiconductor theory of barrier layer" translated from German original: "Halbleitertheorie der Sperrschicht" *Naturwissenschaften* **26**, 843 (1938)

Schottky W. "Title unknown to EFS" *Zeitschrift für Physik* **41**, 570 (1940)

Schubert E. F., Tu L. W., Zydzik G. J., Kopf R. F., Benvenuti A., and Pinto M. R. "Elimination of heterojunction band discontinuities by modulation doping" *Applied Physics Letters,* **60**, 466 (1992)

Sharma B. L. and Purohit R. K., *Semiconductor Heterojunctions* (Pergamon, London, 1974)

Shay J. L., Wagner S., and Phillips J. C. "Heterojunction band discontinuities" *Applied Physics Letters* **28**, 31 (1976)

Swaminathan V. and Macrander A. T. *Materials Aspects of GaAs and InP Based Structures* (Prentice Hall, Englewood Cliffs, 1991)

Tersoff J. "Schottky Barrier Heights and the Continuum of Gap States" *Physical Review Letters* **52**, 465 (1984)

Tersoff J. "Schottky barriers and semiconductor band structures" *Physical Review B***32**, 6968 (1985)

Tersoff J. "Summary Abstract: Failure of the common anion rule for lattice-matched heterojunctions" *Journal of Vacuum Science and Technology B***4**, 1066 (1986)

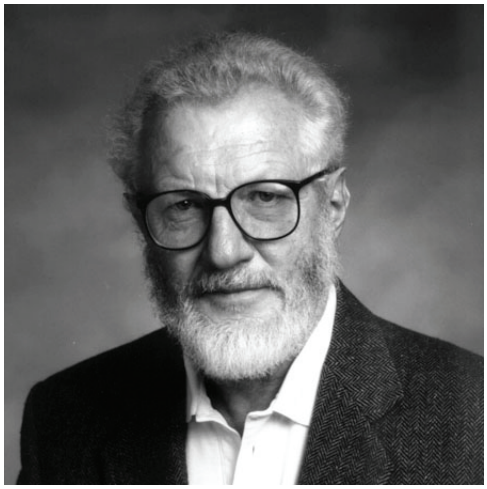
Tien P. K. (1985) unpublished

Tiwari S. and Frank D. J. "Empirical fit to band discontinuities and barrier heights in III-V alloy systems" *Applied Physics Letters* **60,** 631 (1992)

van de Walle C. G. "Band lineups and deformation potentials in the model-solid theory" *Physical Review B***39**, 1871 (1989)

van de Walle C. G. and Martin R. M. "Theoretical calculations of heterojunction discontinuities in the Si/ Ge system" *Physical Review B***34**, 5621 (1986)

Weber J. and Alonso M. I. "Near-band-gap photoluminescence of Si-Ge alloys" *Physical Review B* **40**, 5683 (1989)



Herbert Kroemer (1928)
Pioneer of semiconductor heterostructures

18

Tunneling structures

18.1 Tunneling in ohmic contact structures

Ohmic contacts are non-rectifying metal-semiconductor contacts. They are a key part of any solid-state device. By definition, ohmic contacts exhibit a ***linear*** current-voltage characteristic and the associated resistance is called ***contact resistance***. Ohmic contacts need to be distinguished from ***Schottky contacts*** with are rectifying metal-semiconductor contacts.

Figure 18.1 compares the band diagram of a Schottky contact and an ohmic contact. Ohmic contacts are generally fabricated by very highly doping the semiconductor close to the metal-semiconductor interface. As a result, the depletion region becomes so thin that the barrier can be tunneled through with high probability and thus little resistance.

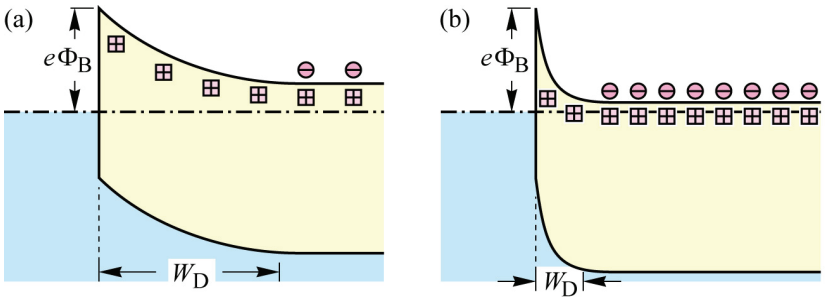


Fig. 18.1. (a) Rectifying Schottky contact and (b) ohmic contact whose high doping results in a thin depletion layer, i.e. a very thin tunneling barrier.

Different fabrication procedures are employed to attain a high doping concentration in the semiconductor and three of these fabrication procedures are shown in ***Fig***. 18.2.

Firstly, a high doping concentration in the semiconductor can be attained by alloying a metal contact with the semiconductor with the metal contact containing an impurity that acts as a donor or acceptor in the semiconductor. For example, AuZn/Au contacts are frequently employed for contacts to p-type GaAs (and other III–V arsenides and phosphides) with the Zn acting as an acceptor impurity. Another example is the AuGe/Ni/Au contact to n-type GaAs (and other III–V arsenides and phosphides) with the Ge acting as a donor impurity.

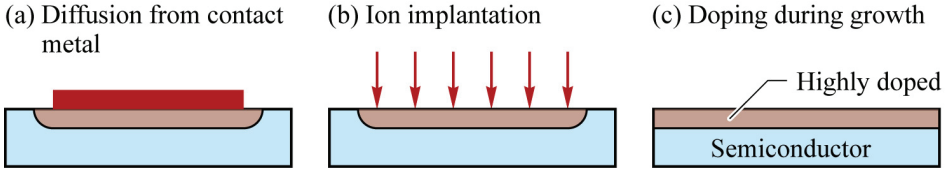


Fig. 18.2. Methods of achieving high doping for ohmic contact formation.

Secondly, a high doping concentration in the semiconductor can be attained by ion-implanting the semiconductor with a high concentration of impurities.

Thirdly, a high doping concentration in the semiconductor can be attained by highly doping the semiconductor with a high concentration of dopants during epitaxial growth.

Although the above discussion of the fabrication of ohmic contacts is helpful and useful, the actual formation of ohmic contacts can be much more complicated: Due to the high temperatures that are used during the alloying process, the contact metal can chemically react with the semiconductor and form new chemical *phases* of material (a "*phase*" is a physically distinct and separable portion of matter). For example, in Si technology, the alloying of the contact metal with the semiconductor results in the formation of metal-silicides, thereby going much beyond a simple doping process of the semiconductor.

An alternative method to form ohmic contacts is the use of a type of metal with a work function that results in a metal-semiconductor barrier height $e\Phi_B$ to be very low so that carriers can easily overcome this barrier by thermionic emission.

18.2 Tunneling current through a triangular barrier

We will next calculate the tunneling current through differently shaped barriers. An illustration of a rectangularly shaped barrier at different bias voltages is shown in ***Fig***. 18.3 (Simmons, 1963). The tunneling probability of a single electron can be calculated from the WKB approximation which states

$$T = e^{-\int_{x=0}^{L_B} 2\hbar^{-1}\sqrt{2m_B^*\left[U(x)-E\right]}\,dx} \tag{18.1}$$

where m_B* is the effective mass in the barrier material to be tunneled through.

At high bias voltages, the tunneling barrier becomes triangularly shaped, as shown in ***Fig***. 18.3 (c). What is the tunneling current through the triangular barrier? Next, this will be calculated using the formalism given by Simmons (1963) and van der Ziel (1976).

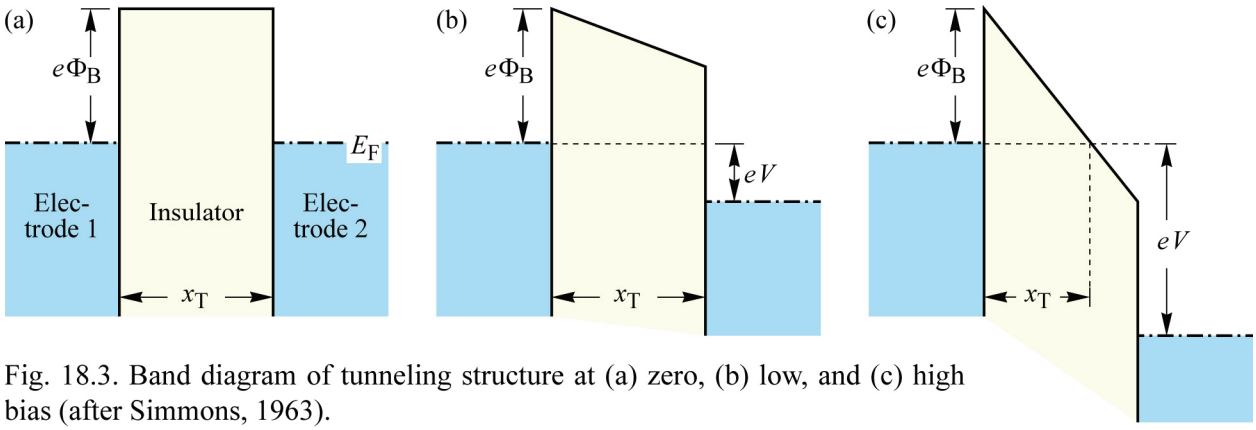


Fig. 18.3. Band diagram of tunneling structure at (a) zero, (b) low, and (c) high bias (after Simmons, 1963).

The band diagram of an emitter electrode (e.g. a semiconductor) and a triangular barrier (e.g. an oxide) is shown in ***Fig***. 18.4. The structure has the potential energy

$$U(x) = 0 \qquad \text{for } x < 0 \tag{18.2a}$$

$$U(x) = e\Phi_B - e\mathcal{E}\,x \qquad \text{for } x \geq 0 \tag{18.2a}$$

where $\mathcal{E}$ is the electric field in the tunneling barrier.

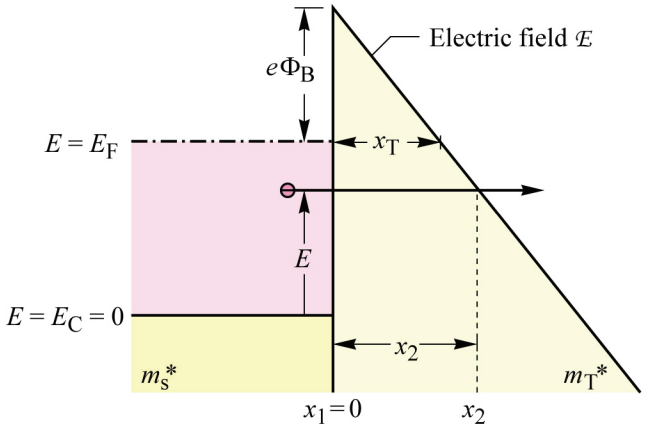


Fig. 18.4. Band diagram of triangular tunneling structure defined by barrier height Φ_B and electric field $\mathcal{E}$.

$x_T = \Phi_B/\mathcal{E}$ (minimum tunnel distance)

m_s^* (electron effective mass in semiconductor)

m_T^* (electron effective mass in tunnel medium)

An electron with energy E, tunneling from the semiconductor through the barrier enters the barrier at $x_1 = 0$ and exits the barrier at $x_2 = (E_F + e\Phi_B - E)/(e\mathcal{E})$. Thus the exponent in Eq. (18.1) becomes

$$-2\frac{\sqrt{2m_B^*}}{\hbar}\int_{x=0}^{x_2}\sqrt{U(x)-E}\;\mathrm{d}x \;=\; -2\frac{\sqrt{2m_B^*}}{\hbar}\;\frac{2}{3}\;\frac{(E_F+e\Phi_B-E)^{3/2}}{e\mathcal{E}} \tag{18.3}$$

Giving a transmission coefficient of

$$T \;=\; \exp\left(-\frac{4}{3}\;\frac{\sqrt{2m_B^*}}{\hbar}\;\frac{(E_F+e\Phi_B-E)^{3/2}}{e\mathcal{E}}\right) \tag{18.4}$$

The energy E is the energy of electrons, and only those electrons are relevant that are incident on the barrier, i.e. those electrons with a positive v_x component. Therefore $E = ½\, m_s^*\, v_x^2 = p_x^2/(2\,m_s^*)$, where m_s^* is the electron effective mass in the semiconductor and p_x is the carrier momentum along the positive x direction.

We now make use of the fact that one electron occupy the phase space "volume" $\Delta x\,\Delta y\,\Delta z\,\Delta p_x\,\Delta p_y\,\Delta p_z = h^3/2$, so that the concentration of electrons (per unit volume) with momentum between p_x and $(p_x + \Delta p_x)$, p_y and $(p_y + \Delta p_y)$, and p_z and $(p_z + \Delta p_z)$ is given by $(2/h^3)\,\mathrm{d}p_x\,\mathrm{d}p_y\,\mathrm{d}p_z$. Thus the number of arriving electrons at the barrier surface per unit time per unit area is given by

$$v_x\;\frac{2}{h^3}\;\mathrm{d}p_x\,\mathrm{d}p_y\,\mathrm{d}p_z \qquad\qquad \text{for } v_x > 0 \tag{18.5}$$

The associated current density of tunneling electrons is then obtained by integration over all momenta and by multiplication with the tunneling probability, i.e.

$$J \;=\; \frac{2e}{m_s^*\,h^3}\int_{p_z}\int_{p_y}\int_{p_x} p_x\,T(p_x)\,\mathrm{d}p_x\,\mathrm{d}p_y\,\mathrm{d}p_z\;. \tag{18.6}$$

The integration must be carried out for all electrons in the conduction band, that is, over all momenta. At the Fermi energy, the electron momentum is given by $p_F = (2\,m_s^*\,E_F)^{1/2}$. Thus the momentum satisfies the condition

$$\mathrm{d}p_x^2 + \mathrm{d}p_y^2 + \mathrm{d}p_z^2 \quad \le \quad 2m_s^* E_F \quad = \quad p_F^2 \ . \tag{18.7}$$

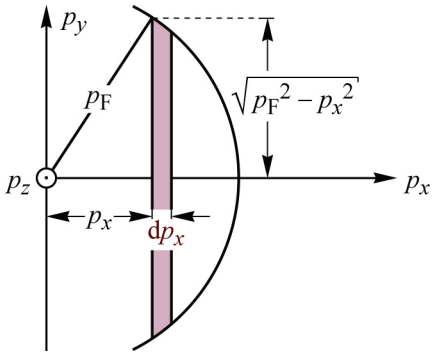


Fig. 18.5. Illustration proofing the following relationship:

$$\int\int \mathrm{d}p_y\, \mathrm{d}p_z = \pi\,(\mathrm{d}p_F^2 - \mathrm{d}p_x^2)$$

To integrate with respect to p_y and p_z, we turn to ***Fig***. 18.5 which shows that

$$\int_{p_z} \int_{p_y} \mathrm{d}p_y\, \mathrm{d}p_z \quad = \quad \pi\left(p_F^2 - p_x^2\right). \tag{18.8}$$

Therefore, Eq. (18.6) can be written as

$$J \quad = \quad \frac{2\pi e}{m_s^*\, h^3} \int_0^{p_F} p_x\, T(p_x) \left(p_F^2 - p_x^2\right) \mathrm{d}p_x \ . \tag{18.9}$$

Realizing that $T(p_x)$ is greatest for $p_x \approx p_F$ and decreases rapidly as p_x decreases, we introduce the new variable $\theta = p_F - p_x$ so that the integrant will have appreciable values only near $\theta \approx 0$. We may thus write, with good approximation

$$p_x \ \approx \ \theta \tag{18.10a}$$

$$p_F^2 - p_x^2 \quad = \quad (p_F + p_x)\,(p_F - p_x) \quad \approx \quad 2p_F\,\theta \tag{18.10b}$$

$$\left(E_F + e\Phi_B - E\right)^{3/2} \ = \left(e\Phi_B + \frac{p_F^2 - p_x^2}{2m_s^*}\right)^{3/2} \approx \left(e\Phi_B\right)^{3/2} + \frac{3}{2}\frac{\sqrt{e\Phi_B}}{m_s^*}\, p_F\,\theta \tag{18.10c}$$

where the last of the three equations is found from a Taylor-series expansion in θ that is truncated after the linear term. Integrating Eq. (18.9) over θ and no longer over p_x, the *lower limit* of the integral becomes 0. Since $T(p_F - \theta)$ decreases rapidly with increasing θ, the *upper limit* of the integral of Eq. (18.9) can be extended to ∞. Substituting into Eq. (18.9), yields

$$J \quad = \quad \frac{2\pi e}{m_s^*\, h^3}(-1) \int_0^{\infty} p_F \exp\left(-\frac{4}{3}\frac{\sqrt{2m_B^*}}{\hbar} \ \frac{\left(e\Phi_B\right)^{3/2} + \frac{3}{2}\frac{\sqrt{e\Phi_B}}{m_s^*}\, p_F\,\theta}{e\mathcal{E}} \right) 2p_F\ \theta\ \mathrm{d}\theta$$

$$= \frac{4\pi e\, p_F^2}{m_s^*\, h^3} \exp\left(-\frac{4}{3} \frac{\sqrt{2 m_B^*}}{\hbar} \frac{(e\Phi_B)^{3/2}}{e\mathcal{E}} \right) \int_0^\infty \exp\left(-2 \frac{\sqrt{2 m_B^*}}{\hbar} \frac{\sqrt{e\Phi_B}}{m_s^*\, e\mathcal{E}}\, p_F\, \theta \right) \theta\, d\theta$$

$$= \frac{m_s^*}{m_B^*} \frac{e^3 \mathcal{E}^2}{8\pi h (e\Phi_B)} \exp\left(-\frac{4}{3} \frac{\sqrt{2 m_B^*}}{\hbar} \frac{(e\Phi_B)^{3/2}}{e\mathcal{E}} \right) \tag{18.11}$$

where we have used the following mathematical relationship

$$\int_0^\infty \exp(-a\theta)\, \theta\, d\theta = \frac{1}{a^2} \qquad \text{where} \qquad a = 2 \frac{\sqrt{2 m_B^*}}{\hbar} \frac{\sqrt{e\Phi_B}}{m_s^*\, e\mathcal{E}}\, p_F\,. \tag{18.12}$$

Neglecting the difference in effective mass (i.e. using $m_s^*/m_B^* = 1$), which does not introduce a large error, one obtains

$$\boxed{J = \frac{e^2 \mathcal{E}^2}{8\pi h \Phi_B} \exp\left(-\frac{4}{3} \frac{\sqrt{2 m^*}}{\hbar} \frac{(e\Phi_B)^{3/2}}{e\mathcal{E}} \right)} \tag{18.13}$$

which is known as the ***Fowler–Nordheim tunneling formula***. This formula is very useful when calculating tunneling currents through barriers.

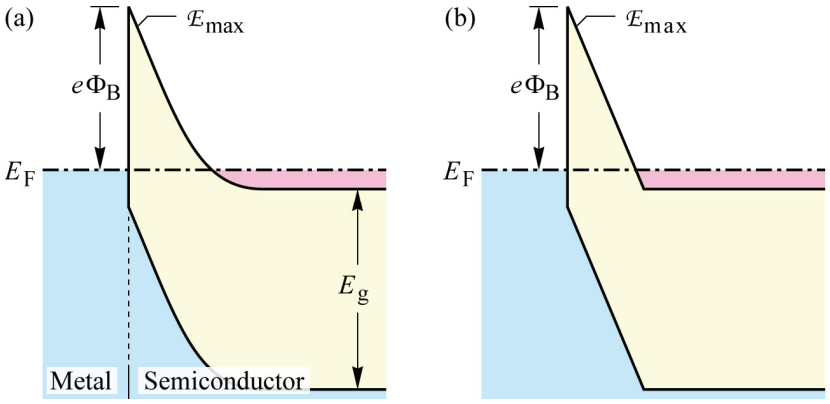


Fig. 18.6. (a) Band diagram of ohmic contact consisting of metal electrode and highly doped semiconductor having a maximum electric field of $\mathcal{E}_{max}$. (b) Band diagram having same maximum electric field, $\mathcal{E}_{max}$, used in the calculation of specific contact resistance.

18.3 Contact resistance of highly doped ohmic contact

We will next calculate the resistance of a highly doped ohmic contact. Consider an n-type ohmic contact, i.e. a metal-semiconductor junction with a semiconductor, highly doped with donors of concentration N_D. The band diagram of such an ohmic contact is shown in ***Fig***. 18.6 (a). The depletion width, W_D, and maximum electric field in the semiconductor, $\mathcal{E}_{max}$, which occurs at the metal-semiconductor boundary, are given by

$$W_D = \sqrt{\frac{2\varepsilon}{e N_D} \Phi_B} \qquad \mathcal{E}_{max} = \sqrt{\frac{2 e N_D}{\varepsilon} \Phi_B} \tag{18.14}$$

Next, we approximate the barrier by a triangular barrier with the same electric field, as shown in ***Fig***. 18.6 (b). Since the thickness of this barrier is given by $W_B = \Phi_B / \mathcal{E}_{max}$, an additional voltage drop across the contact will result in an electric field given by

$$\mathcal{E}(V) = \mathcal{E}_{max} + \frac{V}{W_B} = \mathcal{E}_{max} + \frac{V}{\Phi_B}\mathcal{E}_{max} = \mathcal{E}_{max}\left(1+\frac{V}{\Phi_B}\right) = \sqrt{\frac{2eN_D}{\varepsilon}\Phi_B}\left(1+\frac{V}{\Phi_B}\right) \quad (18.15)$$

Insertion of the electric field into Eq. (18.13) yields

$$\begin{aligned} J &= \frac{e^2\,\mathcal{E}(V)^2}{8\pi h \Phi_B}\exp\left(-\frac{4}{3}\,\frac{\sqrt{2m^*}}{\hbar}\,\frac{(e\Phi_B)^{3/2}}{e\,\mathcal{E}(V)}\right) \\ &= \frac{e^2\left[\sqrt{\frac{2eN_D}{\varepsilon}\Phi_B}\left(1+\frac{V}{\Phi_B}\right)\right]^2}{8\pi h \Phi_B}\exp\left(-\frac{4}{3}\,\frac{\sqrt{2m^*}}{\hbar}\,\frac{(e\Phi_B)^{3/2}}{e\sqrt{\frac{2eN_D}{\varepsilon}\Phi_B}\left(1+\frac{V}{\Phi_B}\right)}\right) \end{aligned} \quad (18.16)$$

This is the current-voltage characteristic of an ohmic contact. Next, we differentiate J with respect to V to obtain the contact resistance. We use

$$\frac{dJ}{dV} = \frac{dJ}{d\mathcal{E}(V)}\,\frac{d\mathcal{E}(V)}{dV} \quad (18.17)$$

where

$$\begin{aligned} \frac{dJ}{d\mathcal{E}(V)} = \frac{e^2}{8\pi h\Phi_B}\Bigg\{ & 2\mathcal{E}(V)\exp\left(-\frac{4}{3}\,\frac{\sqrt{2m^*}}{\hbar}\,\frac{(e\Phi_B)^{3/2}}{e\,\mathcal{E}(V)}\right) + \\ & + \mathcal{E}(V)^2\exp\left(-\frac{4}{3}\frac{\sqrt{2m^*}}{\hbar}\frac{(e\Phi_B)^{3/2}}{e\,\mathcal{E}(V)}\right)\times\left[-\frac{4}{3}\frac{\sqrt{2m^*}}{\hbar}\frac{(e\Phi_B)^{3/2}}{e}\left(\frac{-1}{\mathcal{E}(V)^2}\right)\right]\Bigg\} \end{aligned} \quad (18.18)$$

and

$$\frac{d\mathcal{E}(V)}{dV} = \sqrt{\frac{2eN_D}{\varepsilon}\Phi_B}\,\frac{1}{\Phi_B} = \sqrt{\frac{2eN_D}{\varepsilon\Phi_B}}\,. \quad (18.19)$$

The specific contact resistance, ρ_c, is then given by

$$\rho_c = \left(\frac{dJ}{dV}\right)^{-1}\Bigg|_{V=0} = \left(\frac{dJ}{d\mathcal{E}(V)}\,\frac{d\mathcal{E}(V)}{dV}\right)^{-1}\Bigg|_{V=0} \quad (18.20)$$

As an example, we consider the ohmic contact resistance of n-type GaAs as a function of the doping concentration. Using $\varepsilon_r = 13.1$, $\Phi_B = 0.9$ V, and $m^* = 0.067\,m_0$, we calculate the specific contact resistance as a function of the donor concentration. The result is shown in ***Fig***. 18.7.

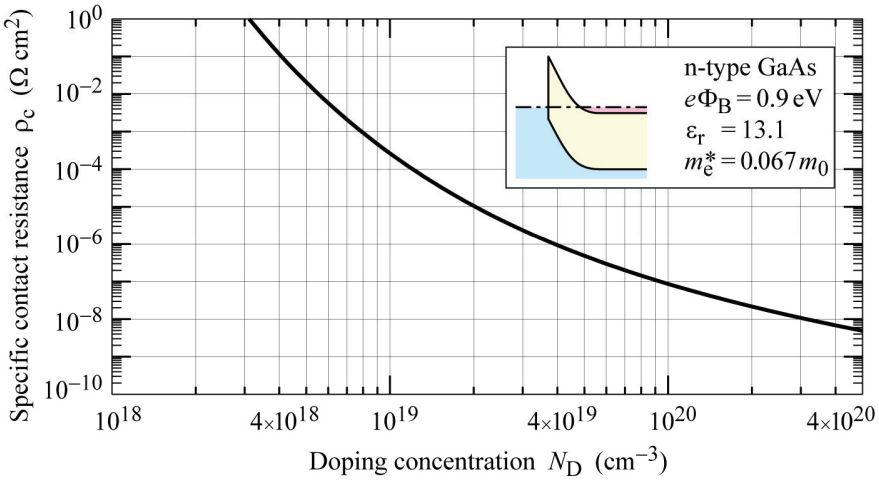


Fig. 18.7. Calculated specific contact resistance of ohmic contact to n-type GaAs as a function of doping concentration.

18.4 Resonant-tunneling structures

Resonant tunneling structures (RTS) consist of two tunneling barriers and a quantum well layer between the barriers. Current flows from the one electrode, denoted as *emitter*, through the double barrier structure to the receiving electrode, denoted as *collector*, as shown in ***Fig***. 18.8. Emitter and collector are doped to provide charge carriers for transport. RT structures are usually n-type due to the larger quantum energies attainable with the lighter electrons as compared to the heavier holes.

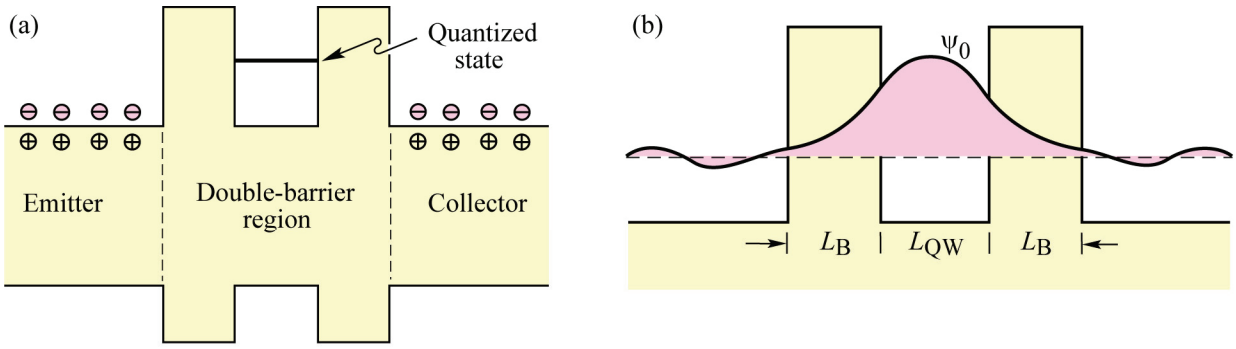


Fig. 18.8. (a) Band diagram of an n-type resonant-tuneling structure and (b) ground-state wave function in the well.

The quantum well layer is sufficiently thin so that just one or two quantized states occur in the well. The tunneling barriers are sufficiently this so that the tunneling probability through one of the barriers is sufficiently large to carry a current density of reasonable magnitude, *e. g.* in the A / cm^2 to kA / cm^2 range. As a bias is applied to the RTS, the energy of quantum state in the center well will change, and, at some bias, the quantum state will be in resonance with electrons injected from the emitter.

Let us first assume that the RTS is biased in such a way that the emitter is in resonance with the quantum state in the quantum well. In this case, carriers need to tunnel only through the first barrier to reach an allowed state. The tunneling probability can be conveniently calculated by the WKB approximation. Recall that the tunneling probability is given by

$$T = e^{-\int_{x=0}^{L_B} 2\hbar^{-1}\sqrt{2m^*[U(x)-E]}\,dx} \tag{18.21}$$

where L_B is the thickness of the tunnel barrier. *All* carriers that tunnel through the first barrier and reach the well, will eventually escape from the well and tunnel through the second barrier to the lower-energy states of the collector.

Let us next consider the case that the RTS is biased in such a way that the emitter is *not* in resonance with the quantum state in the quantum well. In this case, carriers need to tunnel through the first barrier, the well, and the second barrier to reach an allowed state in the collector. Again, the tunneling probability can be conveniently calculated by the WKB approximation. However, the carriers need to tunnel much farther.

$$T = e^{-\int_{x=0}^{L_B+L_W+L_B} 2\hbar^{-1}\sqrt{2m^*[U(x)-E]}\,dx} \tag{18.22}$$

Comparison of the tunneling probabilities for the on-resonance case (Eq. 18.21) with the off-resonance case (Eq. 18.22) yields that the tunneling probability is different by many orders of magnitude. There is a distinct peak in the current-voltage characteristic is expected when the resonance condition is satisfied.

The current-voltage (*I-V*) curve of a resonant tunneling structure is shown in ***Fig***. 18.9 for different bias conditions. The *I-V* curve exhibits a clear peak and decreases again at higher bias voltages. If an RTS has several levels in the well, several peaks in the *I-V* characteristic might be observed. However, if the well has several levels the energy spacing between them is reduced and any broadening mechanism will obscure the manifestation of distinct peaks.

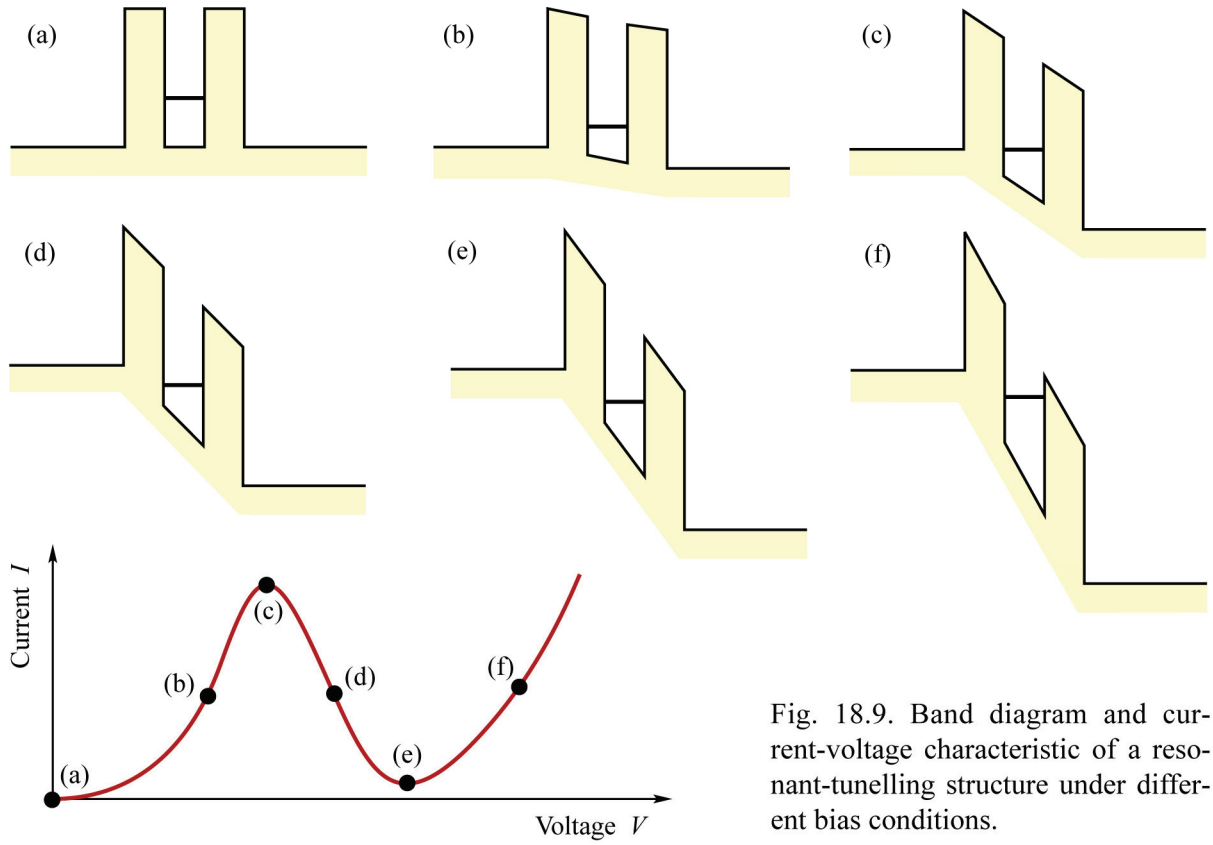


Fig. 18.9. Band diagram and current-voltage characteristic of a resonant-tunelling structure under different bias conditions.

Potential drops in a resonant tunneling structure

There are different contributions to the potential drop in a resonant tunneling structure. Assume that the RTS is under bias and a potential drop occurs at the resistive region, *i. e*. the double barrier region. The different potential drops are shown in ***Fig***. 18.9.

As the bias is applied to the structure an accumulation layer forms at the emitter of the RTS. Band bending is induced by free carriers accumulating in front of the first barrier. At the collector side, the band bending is induced by a depletion region, *i. e*. by fixed donor charges. In the center well region, a small stored charge will be stored if a current flows across the double barrier structure. At all boundaries, the electrostatic boundary condition must be fulfilled, namely that the normal component of the electric displacement $D = \varepsilon \mathcal{E}$ be continuous.

The total potential drop is a sum of the potential drops in the accumulation layer, the barrier and well region, and the depletion region. We assume that the electric field in the first barrier is given by $\mathcal{E}$. We next calculate the different potential drops in the RTS.

The electric field at the emitter-barrier layer interface in the emitter layer is given by the boundary condition $\varepsilon_{Emit}\, \mathcal{E}_{Emit} = \varepsilon_B\, \mathcal{E}$ due to the boundary condition. For simplicity, we assume that the semiconductors forming the RTS have the same dielectric constant so that $\varepsilon_{Emit} = \varepsilon_B$ and thus $\mathcal{E}_{Emit} = \mathcal{E}_B$, *i. e*. the electric field is continuous at the boundaries. The potential in the accumulation layer of the emitter is shown in ***Fig***. 18.10. The band bending in the accumulation layer is caused by the electrons in the accumulation layer. Thus the total charge in the accumulation ayer is given by Gauss' law

$$\mathcal{E} = \frac{e\, n_{acc}^{2D}}{\varepsilon} \tag{18.23}$$

where n_{acc}^{2D} is the density of electrons per unit area in the accumulation layer.

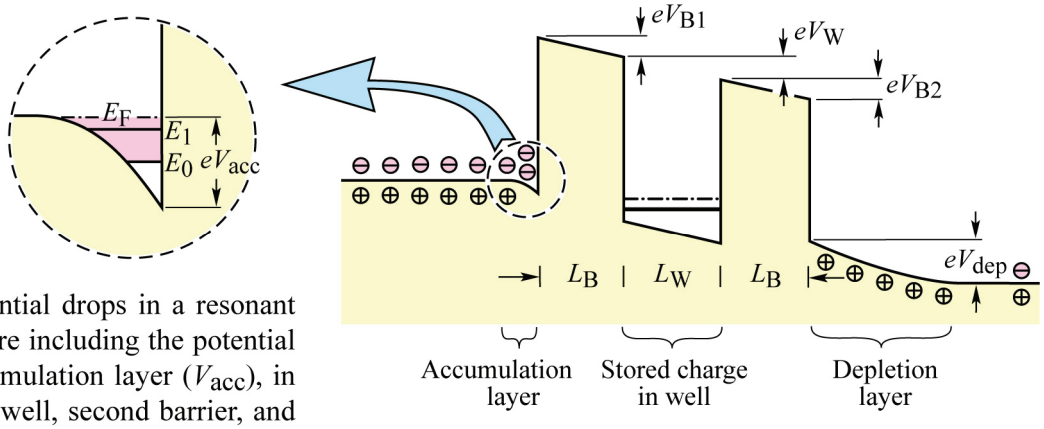


Fig. 18.10. Potential drops in a resonant tunelling structure including the potential drop in the accumulation layer (V_{acc}), in the first barrier, well, second barrier, and the depletion region (V_{dep}).

Due to the narrow size of the triangular accumulation layer potential, size quantization occurs. A suitable wave function for electrons in the accumulation layer is the Fang-Howard wave function (which was discussed earlier). The ground-state energy level is given by

$$E_0 = \frac{3}{2}\left(\frac{3}{2}\frac{e\,\mathcal{E}\,\hbar}{\sqrt{m^*}}\right)^{2/3} \tag{18.24}$$

Assuming that the quantum well has only one quantized state, the Fermi level can be inferred from the electron concentration according to

$$n_{\mathrm{acc}}^{2\mathrm{D}} = \int_{E_0}^{\infty} \rho^{2\mathrm{D}}(E)\, f_{\mathrm{FD}}(E)\, \mathrm{d}E \tag{18.25}$$

where ρ^{2D} is the two-dimensional density of states. In the high-density approximation, this simplifies to

$$n_{\mathrm{acc}}^{2\mathrm{D}} = \rho^{2\mathrm{D}}\,(E_{\mathrm{F}} - E_0)\ . \tag{18.26}$$

Thus the accumulation potential is given by

$$eV_{\mathrm{acc}} = E_0 + (E_{\mathrm{F}} - E_0) = \frac{3}{2}\left(\frac{3}{2}\frac{e\mathcal{E}\hbar}{\sqrt{m^*}}\right)^{2/3} + \frac{n_{\mathrm{acc}}^{2\mathrm{D}}}{\rho^{2\mathrm{D}}} = \frac{3}{2}\left(\frac{3}{2}\frac{e\mathcal{E}\hbar}{\sqrt{m^*}}\right)^{2/3} + \frac{\varepsilon\mathcal{E}}{e\rho^{2\mathrm{D}}} \tag{18.27}$$

We have thus expressed the voltage drop in the accumulation layer as a function of the electric field.

Next we determine the potential drop in the center regions of the RTS. We first assume, that the charge stored in the well is negligibly small. Then, the potential drops in the first barrier, the well, and the second barrier are simply given by

$$V_{\mathrm{B1}} = \mathcal{E}\, L_{\mathrm{B1}}, \qquad V_{\mathrm{W}} = \mathcal{E}\, L_{\mathrm{W}}, \text{ and} \qquad V_{\mathrm{B2}} = \mathcal{E}\, L_{\mathrm{B2}}\ . \tag{18.28}$$

If the charge in the well is taken into account, the electric field in the second barrier is further increased by that charge. The increase can be calculated from Gauss' law. The charge in the center well will be discussed in detail below.

The potential drop in the collector layer is caused by the charged depletion layer. If the collector is doped with a donor concentration N_{D}, and the electric field in the collector at the barrier-collector boundary is given by $\mathcal{E}$, then Gauss' law yields the thickness of the depletion layer

$$W_{\mathrm{D}} = \varepsilon\,\mathcal{E} / N_{\mathrm{D}} \tag{18.29}$$

Using Poisson's equation, one obtains the potential drop in the depletion layer

$$V_{\mathrm{dep}} = \frac{e}{2\,\varepsilon} N_{\mathrm{D}}\, {W_{\mathrm{D}}}^2\ . \tag{18.30}$$

The total voltage applied to the device is the sum of the different potentials discussed above, *i. e.*

$$V = V_{\mathrm{acc}} + V_{\mathrm{B1}} + V_{\mathrm{W}} + V_{\mathrm{B2}} + V_{\mathrm{dep}}\ . \tag{18.31}$$

We thus can determine the relationship between the electric field in the RTDs and the applied voltage.

The attempt to escape model

Assume that an electron has tunneled through the emitter barrier and is confined by the quantum

well. The electron confined by the well has a kinetic energy and thus bounces back and forth between the two barriers. Each time the electron impinges on the barrier, it attempts to escape from the well. However, due to the low tunneling probability, the electron is unlikely to escape from the well at its first attempt. After sufficiently many attempts, the electron will eventually escape from the well. It will escape through the second barrier into the collector due to the availability of unoccupied states in the collector. The escape to the emitter is unlikely because electrons in the accumulation layer occupy states with the same energy as the electrons in the well.

The rate of attempts to escape from the quantum well can be derived from the kinetic energy of electrons in the well. Assume that electrons in the well have a state energy E_0 with respect to the bottom of the well. Using the infinite well-approximation, the energy and the k value of the electron are related by

$$E_0 = \frac{\hbar^2}{2\,m^*}\left(\frac{\pi}{L_{QW}}\right)^2 = \frac{\hbar^2\,k^2}{2\,m^*} = \frac{p^2}{2\,m^*} . \tag{18.32}$$

Thus the electron velocity is give by

$$v = \frac{p}{m^*} = \sqrt{\frac{2\,E_0}{m^*}} . \tag{18.33}$$

The attempt rate of the electron to escape through the exit barrier is then given by

$$A = \frac{v}{2\,L_{QW}} = \frac{\sqrt{2\,E_0/m^*}}{2\,L_{QW}} = \frac{\sqrt{2\,E_0/m^*}}{2\,\pi\,\hbar/\sqrt{2\,E_0\,m^*}} = \frac{2\,E_0}{\pi\,\hbar} . \tag{18.34}$$

where we have used the infinite well approximation to express L_{QW} in terms of E_0. Each attempt has a success probability of T, the tunneling probability. Thus the rate of successful attempts is given by

$$AT = \frac{2\,E_0}{\pi\,\hbar}\,T . \tag{18.35}$$

The inverse of the rate of successful attempts is the lifetime of the electron in the well, that is,

$$\Delta\tau = (A\,T)^{-1} = \frac{\pi\,\hbar}{2\,E_0\,T} . \tag{18.36}$$

The finite lifetime of electrons in the well leads, according to the uncertainty principle, to a broadening of the quantum state in the well. Since the uncertainty principle is given by $\Delta E\,\Delta\tau \approx \hbar$, the quantum state has the width

$$\Delta E = \frac{\hbar}{\Delta\tau} = \frac{2\,E_0\,T}{\pi} . \tag{18.37}$$

The broadening of the energy level is schematically shown in ***Fig***. 18.11. The broadening of

the energy level leads to a broadening of the voltage resonance. The width of the voltage resonance is then given by $\Delta V = \Delta E / e$. Calculating the width of the energy resonance for a typical resonant tunneling structure reveals that the calculated energy resonance broadening is much smaller than the linewidth of the voltage peak of a typical resonant tunneling structure. This is because other broadening mechanisms cause additional broadening of the current peak in the *I-V* characteristic. The additional broadening mechanisms include the thermal energy distribution of carriers in the emitter, any random fluctuations of the quantum barriers or the quantum well width, or any compositional fluctuations of the barrier and well materials.

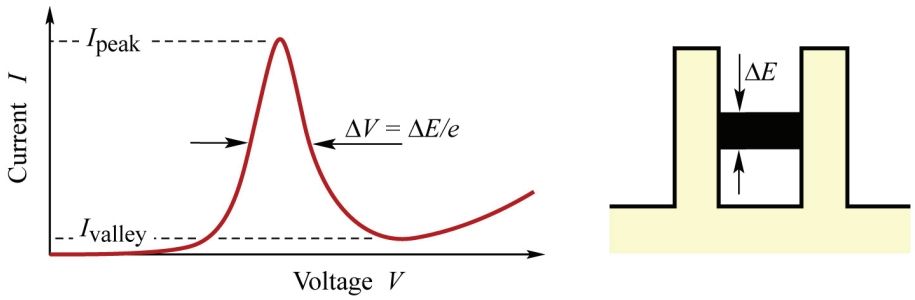


Fig. 18.11. Linewidth of resonance peak. Also shown are the peak and valley current.

Exercise: ***Linewidth of the current resonance in an RTS***. Calculate the linewidth of the current peak in a typical $Al_xGa_{1-x}As$ / GaAs resonant tunneling structure with $x = 30$ %, $L_B = 40$ Å, and $L_{QW} = 100$ Å. Compare your result to typical experimental linewidths in RTS which are several tens of meV at low temperatures and several hundreds meV at room temperature.

Solution:
It is $m^* = 0.067\ m_0$ and thus $E_0 = (\hbar / 2m^*)\ (\pi / L_{QW})^2 = 56.3$ meV. The barrier energy is $U = \Delta E_C = (2/3)\ \Delta E_g = (2/3)\ (1.247 \times 0.3) = 250$ meV. Thus the tunneling probability is given by

$$T = \exp\left[-\int_{x=0}^{L_B} 2\hbar^{-1} \sqrt{2m^* (U_x - E_0)}\ dx \right] = 9.4 \times 10^{-3} .$$

The linewidth is given by

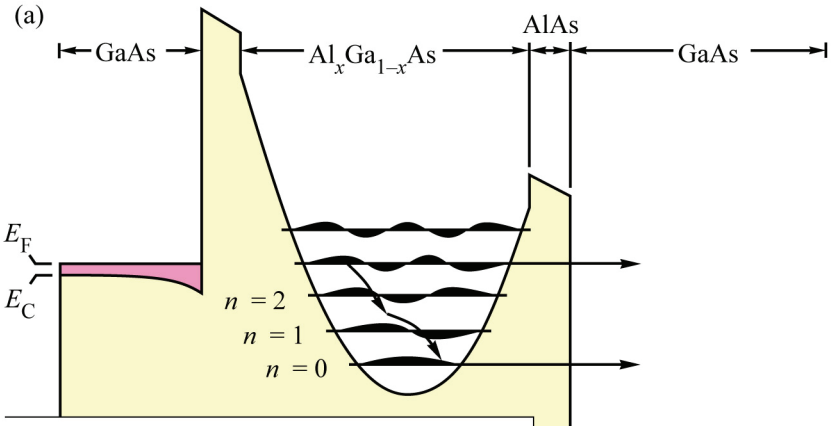
$$\Delta E = (2\, E_0 T / \pi) = 0.338 \text{ meV} .$$

The theoretical linewidth is much smaller than experimental linewidths in RTS. Typical experimental linewidths are several tens of meV at low temperature and several hundreds of meV at room temperature.

Exercise: ***Resonant tunneling structures***. Suggest some ways to improve the peak-to-valley current ratio of resonant tunneling structures. Explain the dependence of the peak-to-valley ratio on the emitter doping concentration and the measurement temperature.

Resonant tunneling structures with a parabolic well

Resonant tunneling structures with a parabolic well have been demonstrated as well. In a parabolic well, the energy levels are equidistant. Such equidistance is indeed confirmed in current-voltage measurements that are shown in ***Fig***. 18.12.



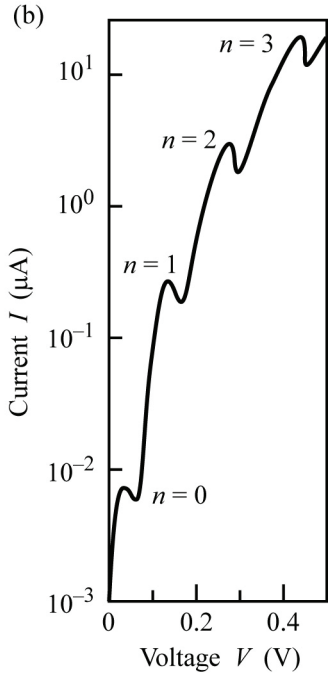


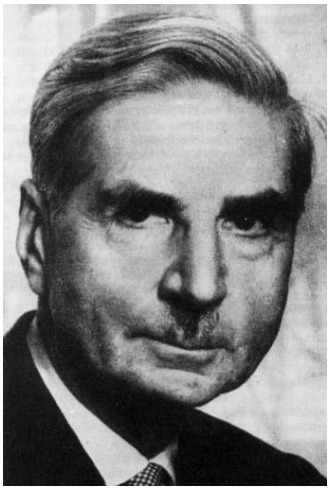
Fig. 18.12. Band diagram of a resonant-tunneling structure with a parabolic well. Energy relaxation in the well is indicated by arrows. (b) Current-voltage characteristic of an $Al_xGa_{1-x}As/GaAs$ parabolic well resonant tunneling structure at T = 4.2 K. Four resonances, n = 0 to n = 3, are observed.

References

Simmons J. G. "Generalized formula for the electric tunnel effect between similar electrodes separated by a thin insulating film" *Journal of Applied Physics* **34**, 1793 (1963)

van der Ziel A. *Solid State Physical Electronics*, 3rd edition (Prentice-Hall, Englewood Cliffs NJ, 1976)

Leo Esaki (1925–)
Developed pn-junction tunnel diodes



Walter Schottky (1886–1976)
Developed theory of metal-semiconductor contacts

19

Electronic transport

This chapter discusses the transport of carriers in a solid-state material, particularly in semiconductors. Carrier motion is ***directed***, when carriers are driven by the force of an electric field. We call such motion of carriers the ***drift*** motion. Carrier motion is ***undirected*** when carriers, driven by their thermal energy, randomly diffuse without a preferential direction. We call such motion of carriers the ***diffusion*** motion. This chapter discusses the carrier transport in semiconductors and other solid-state materials.

19.1 Electrons in an electric field

Let us consider an electron with mass m^* and the negative charge $-e$ being subjected to an electric field $\mathcal{E}$. As a result of the electric field the electron experiences the force

$$F = -e\mathcal{E} \,. \tag{19.1}$$

As a result of the force, the electron is accelerated. By influence of the lattice, the electron is decelerated. This balance of acceleration and deceleration is schematically shown in ***Fig***. 19.1. Thus in an electric field, electrons drift with a constant velocity, i.e.

$$\boxed{v_d = -\mu\mathcal{E}} \tag{19.2}$$

where v_d is the electron ***drift velocity*** and μ is a proportionally constant called the ***electron mobility***.

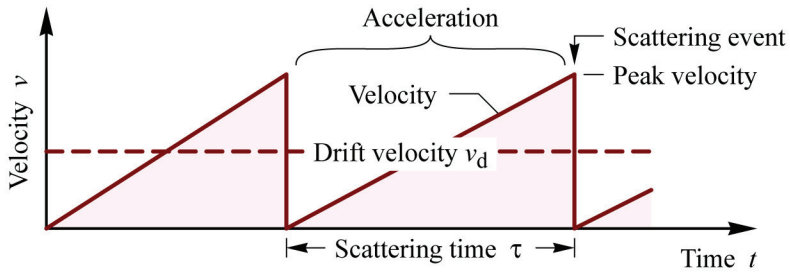


Fig. 19.1. Electron velocity in constant electric field $\mathcal{E}$ that results in acceleration and scattering. Low temperatures are assumed where electron diffusion can be neglected.

We next will evaluate the electron mobility by considering the influence of the electric field on the electron momentum, p, which leads to the *acceleration* of the electron, and the influence of scattering events with the lattice, which leads to a *deceleration* of the electron. We will require that in the steady state, the change in electron momentum due to the accelerating electric field is equal to the change in electron momentum due to the decelerating lattice scattering.

We learned from our quantum-mechanical considerations that a quantum mechanical process has a transition probability per unit time (i.e. the transition *rate*, e.g. from state i to a state j, which was given by Fermi's Golden Rule). The inverse of the transition probability per unit time

is the lifetime of the electron in that state, τ, or briefly the ***scattering time***, τ. Because the electron's momentum is randomized during the scattering event, the time τ is also referred to as the ***momentum-relaxation time***.

The change in electron momentum due to the electron-*accelerating* electric field within the time interval Δt is given by

$$\left.\frac{dp}{dt}\right|_{\text{electric field}} \Delta t \;=\; m^* \frac{dv}{dt}\Delta t \;=\; m^* a\,\Delta t \;=\; F\,\Delta t \;=\; -e\mathcal{E}\,\Delta t \tag{19.3}$$

where a is the acceleration of the electron subjected to an electric field, as shown in ***Fig***. 19.1.

To calculate the *decelerating* rate of change in electron momentum, we consider the probability distribution of a quantum mechanical transition, which is the exponential distribution, as shown in ***Fig***. 19.2.

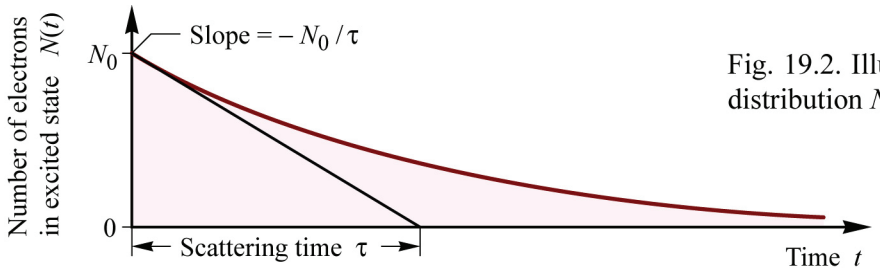


Fig. 19.2. Illustration of exponential distribution $N(t) = N_0 \exp(-t/\tau)$.

Consider a group (or an ensemble) of initially N_0 particles that are subject to a quantum mechanical transition such as a scattering event. The number of un-scattered particles decrease according to $N(t) = N_0 \exp(-t/\tau)$. In the steady state, the number of particles in a given state is a constant, i.e. the particles transitioning out of the state is equal to the number of particles transitioning into that state. The number of particles transitioning out of the state is given by the initial slope of the curve, i.e. by N_0/τ. Thus the change in momentum *of the ensemble of carriers* due to scattering within the time interval Δt is given by

$$\left.\frac{dp}{dt}\right|_{\text{of ensemble due to scattering}} \Delta t \;=\; p\,\frac{dN(t)}{dt}\Delta t \;=\; p\,\frac{N_0}{\tau}\Delta t\ . \tag{19.4}$$

Thus the *average* change in momentum of *one carrier* due to scattering within the time interval Δt is obtained by dividing both sides of the equation by N_0. One obtains

$$\left.\frac{dp}{dt}\right|_{\text{scattering}} \Delta t \;=\; p\,\frac{1}{\tau}\Delta t\ . \tag{19.5}$$

Because we require that in the steady state, the change in electron momentum due to the accelerating electric field is equal to the change in electron momentum due to the decelerating lattice scattering, we equate Eqs. (19.3) and (19.5) and obtain

$$\boxed{\left.\frac{dp}{dt}\right|_{\text{electric field}} \Delta t \;=\; \left.\frac{dp}{dt}\right|_{\text{scattering}} \Delta t} \tag{19.6}$$

or

$$-e\mathcal{E} \quad = \quad p/\tau \ . \tag{19.7}$$

Using that the momentum and the drift velocity are related by $p = m^* v_d$ and solving for the drift velocity yields

$$v_d \quad = \quad -\frac{e\tau}{m^*}\ \mathcal{E} \tag{19.8}$$

Comparing this equation with Eq. (19.2) yields

$$\boxed{\mu \quad = \quad \frac{e\tau}{m^*}} \tag{19.9}$$

where τ is the momentum-relaxation time. The equation shows that carriers are highly mobile, when their mass is light-weight and scattering times are long.

The mobility of carriers is instrumental in determining the maximum speed of operation of a semiconductor device. Undoped Si has an electron mobility of 1500 cm^2/Vs, whereas undoped GaAs has an electron mobility of 8500 cm^2/Vs. This is the primary reason that GaAs is the device with the better high-speed capability than Si. In high-speed multi-GHz applications such as cellular telephony, compound semiconductor devices are found in the high-power transmitter stages.

In addition, a *high mobility* allows one to *reduce parasitic resistances* in semiconductor devices. Low parasitic resistances result in *low-noise operation*, because each resistor is a source of white noise (thermal noise). This is the primary reason that GaAs is the device with the better noise performance than Si. In high-speed multi-GHz applications such as satellite communications and cellular telephony, compound semiconductor devices are found in the low-noise receiver stages. ***Table*** 19.1 gives the electron and hole mobilities in common lightly-doped semiconductors at room temperature.

Material	**Electron mobility** ($cm^2/(Vs)$)	**Hole mobility** ($cm^2/(Vs)$)
GaAs	8000	320
GaN	1800	30
GaP	110	70
InP	5600	150
Si	1360	460
Ge	3900	1900
α-SiC	400	50

Table 19.1: Electron and hole mobilities at room temperature in semiconductors having a low doping concentration.

Exercise: ***Low-field drift and thermal velocity***. Calculate the thermal velocity of electrons in GaAs ($m_e^* = 0.067\ m_e$) at room temperature. Also calculate the drift velocity of electrons in GaAs (μ = 8000 cm^2/Vs) in an electric field of $\mathcal{E}$ = 100 V/cm. What do you conclude from this comparison?

Solution:
Thermal velocity: Using $E_{kin} = ½\, m\, v_{th}^2 = (3/2)\, kT$, one obtains $v_{th} = 4.5 \times 10^7$ cm/s
Drift velocity: Using $v_{drift} = \mu \mathcal{E}$, one obtains $v_{drift} = 8.0 \times 10^5$ cm/s

For low electric fields, the drift velocity is much lower than the thermal velocity. The trajectory of a drifting electron is as shown in ***Fig***. 19.3. Scattering, i.e. the exchange of energy between the electron and the crystal lattice, therefore is mostly determined by the thermal velocity and ***thus can be assumed to occur at regular time intervals***, i.e. at the scattering time, as shown in ***Fig***. 19.1, and does not depend on the electron velocity.

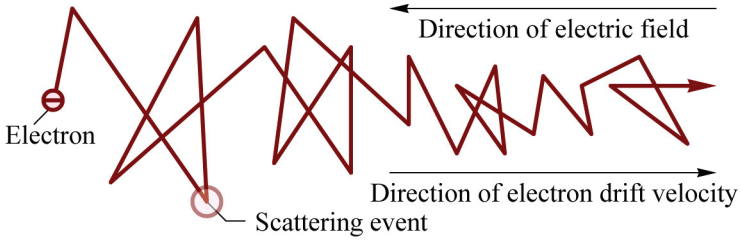


Fig. 19.3. Electron drift and diffusion.

19.2 Matthiessen's rule

Let us assume that we have different carrier-scattering mechanisms, for example the scattering by the crystal lattice (lattice vibrations or phonons) and scattering by ionized impurities. We attribute to each scattering mechanisms a scattering time and a corresponding mobility, that is

Acoustic phonon scattering	μ_{AC}
Optical phonon scattering	μ_{OP}
Ionized impurity scattering	μ_{II}
Neutral impurity scattering	μ_{NI}

What will be the resulting carrier mobility when multiple scattering mechanisms occur? This question was answered by a rule developed by Matthiessen (1906). This rule states

$$\frac{1}{\mu} = \frac{1}{\mu_{AC}} + \frac{1}{\mu_{OP}} + \frac{1}{\mu_{II}} + \frac{1}{\mu_{NI}} + \ldots . \qquad (19.10)$$

Inverse mobilities can be considered the ***hindrances*** that hinder carriers in their motion. Matthiessen's rule implicates that ***hindrances add up***. The validity of the rule is limited to the case when the different scattering mechanisms are independent. We can generalize Matthiessen's rule by writing:

$$\boxed{\frac{1}{\mu} = \sum_i \frac{1}{\mu_i}} \qquad \text{(Matthiessen's rule)} . \qquad (19.11)$$

Exercise: ***Scattering time and mean free path***.
(a) Calculate the scattering time for electrons in GaAs ($m_e^* = 0.067\ m_e$) with a mobility of 8000 cm^2/Vs.
(b) Calculate the mean free path of thermal motion.
(c) Compare the mean free path with the atomic distance and explain result.

Solution:
(a) Scattering time is given by $\tau = \mu\, m_e^* / e = 0.30$ ps
(b) Mean free path = $v_{th} \times \tau = 4.5 \times 10^7$ cm/s × 0.30 ps = 13×10^{-6} cm = 1300 Å
(c) The mean free path is much larger than the inter-atomic distance and the lattice constant, so it takes many periods of the lattice to scatter an electron.

19.3 Scattering mechanisms and low-field mobilities

In a solid-state material such as a semiconductor, there are different scattering mechanisms such as (i) ionized impurity scattering, (ii) neutral impurity scattering, (iii) acoustic phonon scattering, (iv) optical phonon scattering, and (v) alloy scattering. These mechanisms will be discussed below.

Ionized and neutral impurity scattering

Ionized impurity scattering ($\propto 1/\mu_{II}$) is due to the coulombic interaction between a charge carrier and a charged impurity atom. Because the coulombic interaction is very strong, ionized impurity scattering is a most important scattering mechanism in semiconductors.

Neutral impurity scattering, which lacks the strong coulombic interaction of ionized impurity scattering, can be due to the different size, bonding structure, and electronegativity of the impurity atom compared with the regular lattice atom. Because neutral impurity scattering lacks strong coulombic interaction, it is much weaker than ionized impurity scattering. For this reason, we will restrict our considerations to ionized impurity scattering.

Ionized impurity scattering in semiconductors is governed by the same physical principles as Rutherford scattering. In both scattering processes, the trajectory of a charged particle is diverted by the interaction with another charged particle. The coulombic interaction of two charged particles is strongest for small distances between the interacting particles and a long interaction time.

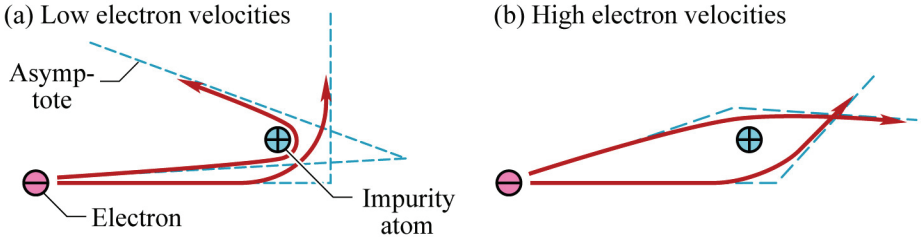


Fig. 19.4. Trajectories of an electron with (a) low and (b) high velocity near an impurity atom. Illustration shows that coulombic interaction is weaker at high electron velocities.

How does ionized impurity scattering depend on temperature? The question can be answered by inspection of ***Fig***. 19.4 which shows electrons propagating on hyperbolic curves, that is, the trajectories are straight lines sufficiently far away from the impurity. At *low* electron velocities, the electron will be close to the impurity for a *long* time and thus the coulombic interaction has a *long* duration, thereby *strongly* deflecting the electron from its path. At *high* electron velocities, the electron will be close to the impurity for a *short* time and thus the coulombic interaction has a *short* duration, thereby *weakly* deflecting the electron from its path. A long interaction time is given for slowly moving electrons, that is, for a non-degenerate carrier gas at low temperatures. The low-temperature mobility is therefore a measure of the impurity and defect content in semiconductors.

Exercise: ***Trajectories of electrons scattered by an ionized impurity***. The analysis of the trajectories of electrons shows that electrons propagate on *hyperbolic curves* when scattered by an impurity. A characteristic of hyperbolic curves is that they have two straight-line asymptotes. Give another example as to where in nature hyperbolic curves occur!

Solution: One example is the hyperbolic trajectory of an outer-space comet that approaches the earth, interacts with the earth's gravitational force, and the leaves the earth's gravitational sphere of influence.

Another example is ***Rutherford scattering***, i.e. when a charged atomic particle, e.g. a He nucleus, is accelerated towards a group of atoms and is scattered by the repulsive coulombic force of another nucleus.

The thermal electron velocity increases with temperature as can be deduced from the equation: $E_{\text{kinetic}} = (1/2)\, mv^2 = (3/2)\, kT$. As the velocity of a mobile carrier increases with temperature, ionized impurity scattering becomes less relevant, because at high velocities the duration of coulombic interaction between impurity and carrier is short. Note that in highly doped semiconductors, the *Fermi velocity* rather than the *thermal velocity* is the relevant velocity.

A detailed analysis reveals the dependence of ionized impurity scattering on temperature as

$$\mu_{\text{II}} \quad \propto \quad T^{3/2} \tag{19.12}$$

That is, ionized impurity scattering ($\propto 1 / \mu_{\text{II}}$) decreases with increasing temperature.

Ionized impurity scattering also depends on the concentration of ionized impurities (or dopants) and thus increases as the concentration of impurities increases. Naively, one would assume that ionized impurity scattering is directly proportional to the number of ionized impurities, *i. e.*

$$\mu_{\text{II}} \quad \propto \quad N_{\text{II}}^{-1} \tag{19.13}$$

However, the dependence is weaker due to the effect of screening. The following dependence has been found

$$\mu_{\text{II}} \quad \propto \quad \left(N_{\text{II}}^{-1} \right)^{\alpha} \qquad \text{where } \alpha < 1 \tag{19.14}$$

At a given temperature, the mobility decreases as the doping concentration increases. This can be verified in ***Fig***. 19.5, which shows experimental carrier mobilities in silicon at room temperature.

The mobility due to ionized impurities was calculated by Conwell and Weisskopf (1950) and later by Brooks and Herring (Brooks, 1955). A review on ionized impurity scattering was given by Chattopadhyay and Queisser (1981). While Conwell and Weisskopf (CW) used unscreened Coulomb potentials, *i. e.* $V = e / (4\pi \varepsilon r)$, Brooks and Herring (BH) used screened Coulomb potentials, *i. e.* $V = (e / 4\pi \varepsilon r)\ \mathrm{e}^{-r/r_s}$, where r_s is the screening length.

The ionized impurity mobility in the Conwell–Weisskopf approximation is given by

$$\mu_{\text{CW}} \quad = \quad \frac{128 \sqrt{2\pi}\ \varepsilon^2\ (kT)^{3/2}}{N_{\text{II}}\ e^3\ \sqrt{m}} \left[\ln\left(1 \ + \ \frac{12\, \pi\, \varepsilon\, kT^2}{N_{\text{II}}^{2/3}\ e^4} \right) \right]^{-1} \tag{19.15}$$

where N_{II} is the concentration of ionized impurities.

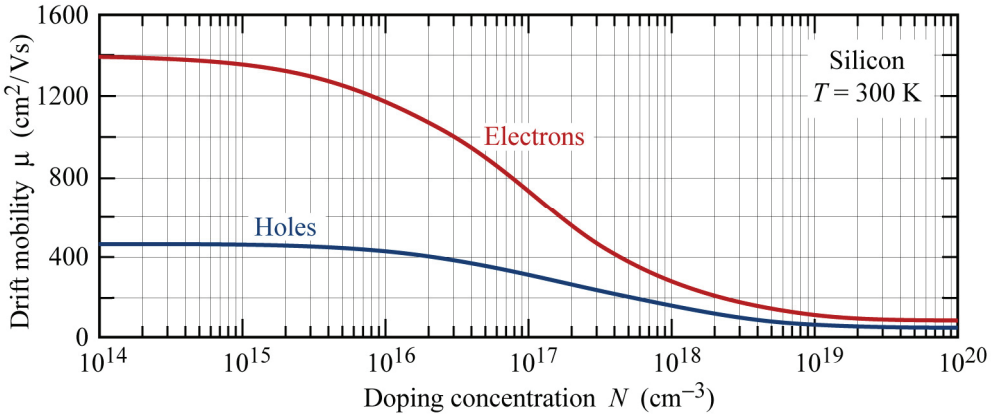


Fig. 19.5. Room temperature electron and hole mobility in silicon versus doping concetration.

The Brooks – Herring approach, which includes the screening of impurities by free carriers, yields for the ionized impurity mobility

$$\mu_{BH} = \frac{128\sqrt{2\pi}\,\varepsilon^2\,(kT)^{2/3}}{N_{II}\,e^3\sqrt{m}}\left[\ln\left(1+\beta_{BH}^2\right) - \frac{\beta_{BH}^2}{1+\beta_{BH}^2}\right]^{-1}. \qquad (19.16)$$

The parameter β_{BH} is given by

$$\beta_{BH} = 2\frac{m}{\hbar}\left(\frac{2}{m}3kT\right)^{1/2} r_D \qquad (19.17)$$

where $r_D = (\varepsilon k T / e^2 n)^{1/2}$ is the Debye screening length, and n is the free carrier concentration. For degenerately doped semiconductors, the Debye screening length must be replaced by the Thomas – Fermi screening length and the thermal energy must be replaced by the Fermi energy.

Both the CW and the BH approach predict a temperature dependence of approximately

$$\mu \propto T^{3/2} \qquad (19.18)$$

i. e. an increasing mobility with temperature. For a non-degenerate semiconductor the experimental mobility indeed approaches zero for $T \to 0$. In *degenerately* doped semiconductors, the Fermi velocity is larger than the thermal velocity at sufficiently low-temperatures and the mobility is expected to remain constant in the low-temperature regime. Note that a qualitative discrepancy exists between the CW and the BH approximation with regard to their density dependence. While the CW-mobility decreases continuously at high concentrations, the BH-mobility first decreases with impurity density but then increases again at very high impurity concentrations (Seeger, 1982). The increase in the BH-mobility is due to screening of the ionized impurity potentials. However, it is not clear if this postulated increase in mobility has ever been observed experimentally.

Phonon scattering

Phonons are quantized lattice vibrations. Consider a one-dimensional (1D) chain of atoms

containing two types of atoms, e.g. the anions and cations in a III–V semiconductor. There are four types of phonons in a 1D chain of atoms, namely the transverse acoustical type (**TA**), the longitudinal acoustical type (**LA**), transverse optical type (**TO**), the longitudinal optical type (**LO**). ***Figure*** 19.6 schematically illustrates the four modes of oscillation in a 1D chain of atoms.

Generally, sound waves propagate in materials by means of LA and TA phonons and they are therefore referred to as "acoustical phonons". Now consider the excitation of atoms by an optical electric field. Due to the different electronegativity of adjacent atoms in a di-atomic lattice, these atoms will move in opposing directions when excited by the optical field, and TO and LO phonons are therefore referred to as "optical phonons".

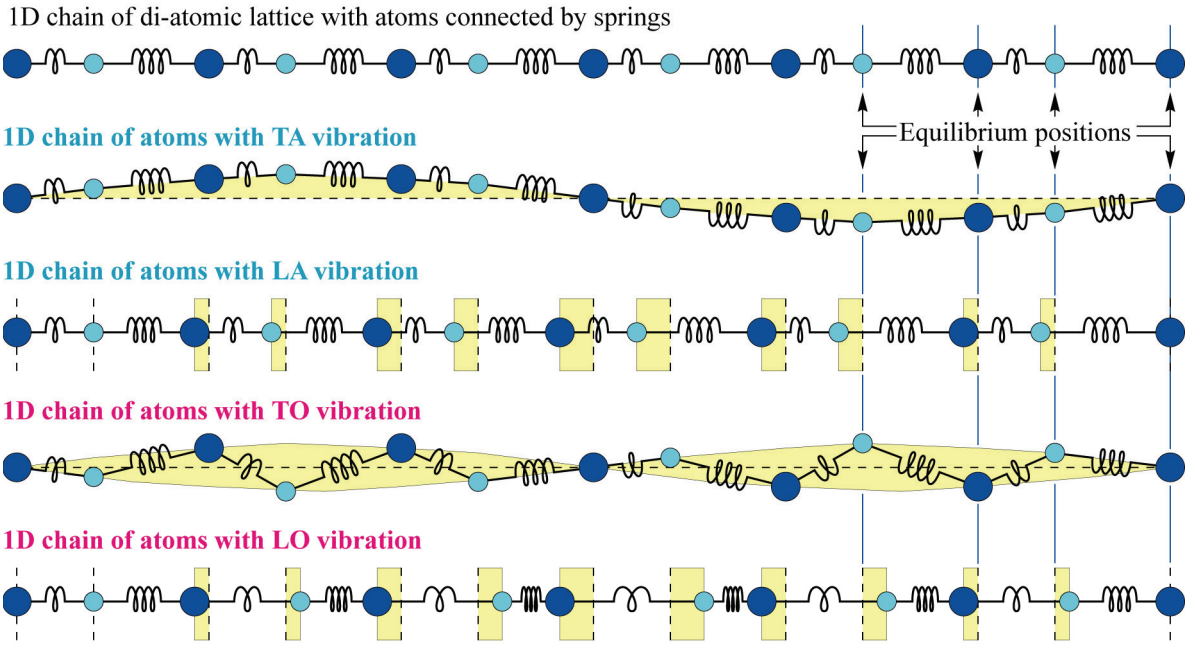


Fig. 19.6. Four vibrational modes of a one-dimensional di-atomic lattice, namely the transverse acoustical (TA), longitudinal acoustical (LA), transverse optical (TO), and longitudinal optical (LO) vibrational mode.

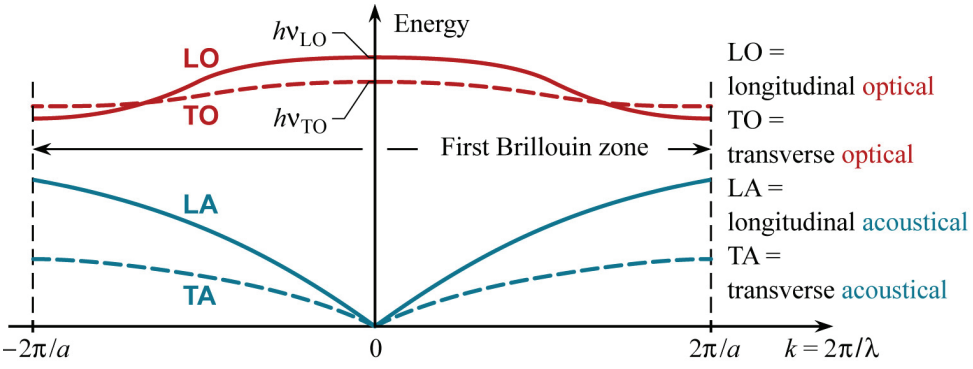


Fig. 19.7. Phonon dispersion relation in di-atomic lattice.

The schematic phonon dispersion relation of a di-atomic 1D lattice is shown in ***Fig***. 19.7. The dispersion relation can be derived using Newtonian mechanics; however, we will not perform the derivation here. Generally the highest energy phonons are LO phonons and they provide the energy loss mechanism of high-energy electrons, e.g. electrons propagating with the

saturation drift velocity. The associated LO phonon emission lifetime is very short, typically 100 fs.

Exercise: ***Phonons***: A phonon is a quantized lattice vibration. Electrons loose or gain energy by emitting or absorbing a phonon, respectively.
(a) What is the direction of motion of adjacent atoms for TA, LA, TO, and LO phonons?
(b) Which type, acoustical or optical type of oscillation has the greater energy?
(c) What is the phase and group velocity of TA and LA phonons?
(d) Is energy of lattice vibrations quantized?
(e) How are phonon frequency and energy related?

Solution:
(a) Adjacent atoms are displaced in the same and opposite directions for acoustical and optical phonons, respectively.
(b) Optical phonons possess the greater energy.
(c) The phase and group velocity can be inferred from the dispersion relation using $v_{phase} = \omega/k$ and $v_{group} = d\omega/dk$.
(d) Yes, the energy of a phonon, just like the energy of any other quantum particle, is quantized.
(e) The energy of acoustical phonons increases as the frequency of oscillation increases. Optical phonons generally have a higher energy due to their inherently higher frequency of oscillation.

Acoustic phonon scattering

Acoustic scattering is due to the emission or absorption of phonons. Acoustic phonon scattering ($\propto 1/\mu_{AC}$) increases with temperature. This is because there are more phonons around at high temperatures compared with low temperatures. A detailed analysis gives the following relation:

$$\mu_{AC} \propto T^{-3/2} \tag{19.19}$$

At any doping concentration, the mobility decreases as temperature increases. At low doping concentrations, the mobility is mostly affected by acoustic phonon scattering. At high doping concentrations, the mobility is affected by both impurity and phonon scattering.

Optical phonon scattering

Optical phonon scattering ($\propto \mu_{OP}$) occurs when carriers have substantial kinetic energy. Longitudinal optical (LO) phonons have an energy of e.g. $h\nu_{LO}$ = 36 meV (for GaAs). To emit an LO phonon, the carrier must have an energy greater or equal to $h\nu_{LO}$.

Experimental results

The electron mobility in n-type silicon versus temperature for different doping concentrations is shown in ***Fig***. 19.8 (Ieong, 2006). Note that for low temperatures, impurity scattering dominates and for $T \to 0$, the mobility approaches zero. Similarly, for very high temperatures (>> 300 K), phonon scattering dominates and for $T \to \infty$, the mobility approaches zero. The calculated electron mobility in moderately doped bulk n-type GaAs is shown in ***Fig***. 19.9 (after Wolfe et al., 1970). In addition to ionized impurity scattering, neutral impurity,

piezoelectric, deformation potential, and polar optical phonon scattering are shown. The combined mobility is also included and is obtained from the sum of the individual scattering mechanisms according to Matthiessen's rule. The mobility has a maximum at a temperature of approximately 40 K.

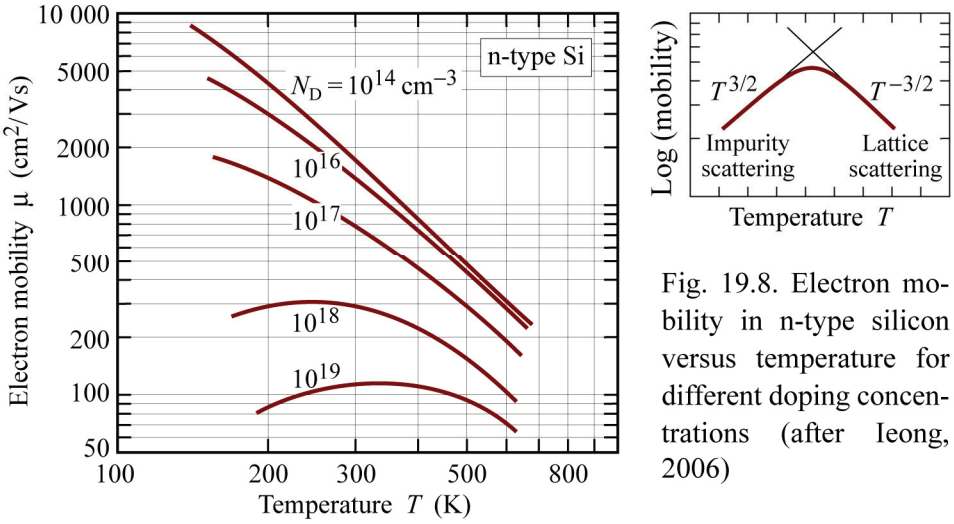


Fig. 19.8. Electron mobility in n-type silicon versus temperature for different doping concentrations (after Ieong, 2006)

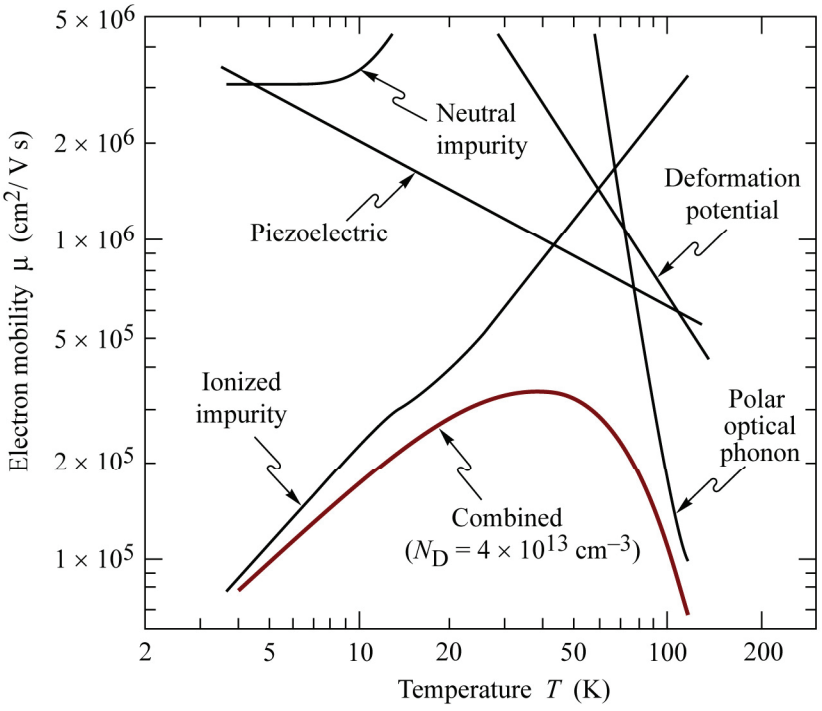


Fig. 19.9. Calculated electron mobilities due to different scattering mechanisms and combined mobility inferred from Matthiessen's rule in high purity GaAs ($N_D = 4 \times 10^{13}$ cm^{-3}) as a function of temperature (after Wolfe et al., 1970).

Alloy scattering

Alloy scattering occurs in all semiconductor random alloys such $Al_xGa_{1-x}As$, $Ga_{1-x}In_xN$, or Si_xGe_{1-x}. Alloy scattering is generally caused by the random distribution of atoms, for example the random distribution of Al and Ga on the cation sites of $Al_xGa_{1-x}As$. By definition, alloy scattering is absent in elemental and binary compound semiconductors such as Si and GaAs.

Consider $Al_xGa_{1-x}N$ at the doping levels $N_D = 10^{17}$ cm^{-3}, 10^{18} cm^{-3}, and 10^{19} cm^{-3}. Let us, for simplicity, assume that the mobility in AlN is about 1/3 of the mobility in GaN. Using the

mobility data shown in the following Section (see future ***Fig***. 19.11), the electron mobilities of the binary compounds are given by

$N_D = 10^{17}$ cm^{-3}	$\mu_{n,GaN} = 900$ cm^2/Vs	$\mu_{n,AlN} = 300$ cm^2/Vs
$N_D = 10^{18}$ cm^{-3}	$\mu_{n,GaN} = 300$ cm^2/Vs	$\mu_{n,AlN} = 100$ cm^2/Vs
$N_D = 10^{19}$ cm^{-3}	$\mu_{n,GaN} = 90$ cm^2/Vs	$\mu_{n,AlN} = 30$ cm^2/Vs

In the *absence* of alloy scattering, the electron mobility in $Al_xGa_{1-x}N$ would just be the linear interpolation between the two mobility values, i.e.

$$\mu_{n,AlGaN} = x\,\mu_{n,AlN} + (1-x)\,\mu_{n,GaN} . \tag{19.20}$$

In the *presence* of alloy scattering, the above mobility and the alloy scattering mobility need to be added in accordance with Matthiessen's rule. The alloy scattering mobility is given by (Look et al., 1992; Morkoc, 1999)

$$\mu_{alloy} = \frac{2^{3/2}\,\pi^{1/2}}{3}\,\frac{e\,\hbar^4}{x\,(1-x)\,V_{cation}\,(e\,V_{alloy})^2\,(m^*)^{5/2}\,\sqrt{kT}} \tag{19.21}$$

where V_{cation} is the volume occupied by one cation and V_{alloy} is the alloy disorder parameter measured in volts. The volume occupied by one cation is given by $V_{cation} = 2/N_{atom}$ where N_{atom} is the total atom concentration (cation plus anion) in GaN. Using the GaN value of $N_{atom} = 8.76 \times 10^{22}$ cm^{-3}, one obtains $V_{cation} = 2.28 \times 10^{-23}$ cm^3.

The ***alloy disorder parameter*** is a critical parameter in determining alloy scattering and the parameter must be related to the potential fluctuations introduced by the random distribution of cations. Although the alloy disorder parameter has been related to the difference in electronegativity of the two cations, Al and Ga, the parameter usually serves as a fitting parameter with typical values of $V_{alloy} = 0.5 - 1.0$ V.

Alloy scattering is proportional to $x(1-x)$, i.e. $\mu_{alloy}^{-1} \propto x(1-x)$, indicating that alloy scattering is absent at $x = 0$ and 1.0 and also indicating that alloy scattering has a maximum at $x = 0.5$.

To calculate the electron mobility in AlGaN, the electron effective mass in AlGaN is needed in Eq. (19.21). Using a linear interpolation, we have

$$m^*_{n,AlGaN} = x\,m^*_{e,AlN} + (1-x)\,m^*_{e,GaN} = x\,0.40\,m_e + (1-x)\,0.20\,m_e . \tag{19.22}$$

The above set of parameters allows us to calculate the electron mobility in AlGaN and the result is shown in ***Fig***. 19.10. Inspection of the figure reveals that alloy scattering introduces a concave bowing that is proportional in magnitude to the alloy disorder parameter.

Summary of scattering mechanisms:

(1) Scattering by phonons occurs at all finite lattice and electron temperatures. Scattering by optical phonons dominates the mobility at temperatures > 300 K.

(2) Neutral and ionized impurity scattering. The scattering by neutral impurities is much weaker than the scattering by ionized impurity atoms due to the lack of Coulomb charge.

(3) Deformation potential and piezoelectric scattering are minor scattering mechanisms (see ***Fig***. 19.9).

(4) Alloy scattering due to random distribution of Al and Ga in the alloy $Al_xGa_{1-x}As$ is a scattering mechanism in all alloy semiconductors. It can be a significant scattering mechanism (see ***Fig***. 19.10).

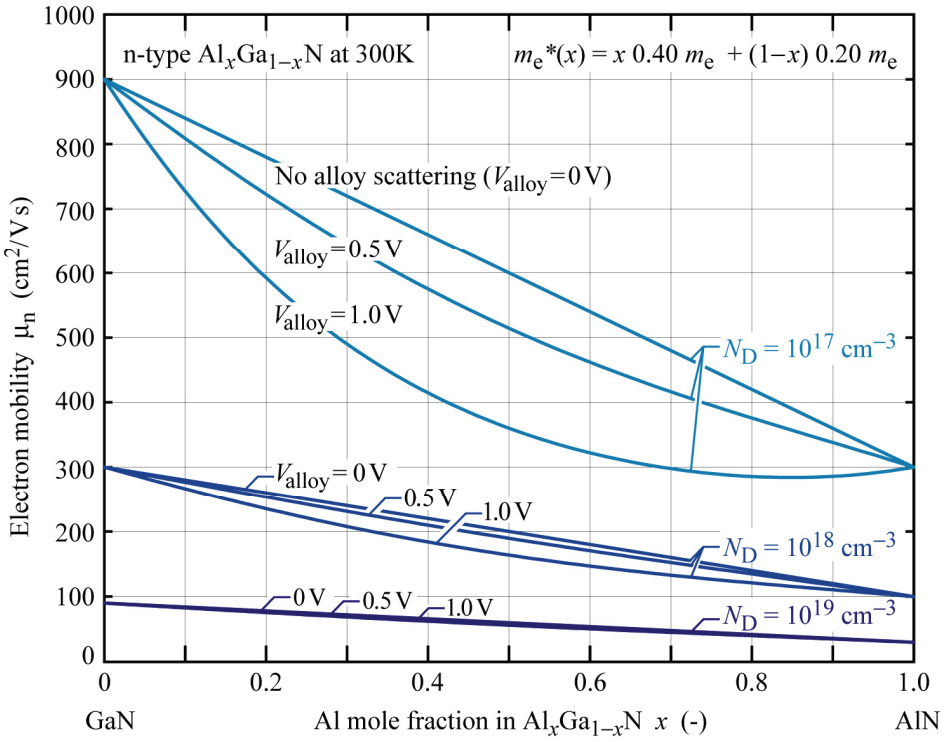


Fig. 19.10. Calculated electron mobility in $Al_xGa_{1-x}N$ as a function of the alloy composition x for different doping concentrations and alloy disorder parameters.

Assumed mobilities:

$N_D = 10^{17}$ cm^{-3}:
$\mu_{GaN} = 900$ cm^2/Vs
$\mu_{AlN} = 300$ cm^2/Vs

$N_D = 10^{18}$ cm^{-3}:
$\mu_{GaN} = 300$ cm^2/Vs
$\mu_{AlN} = 100$ cm^2/Vs

$N_D = 10^{19}$ cm^{-3}:
$\mu_{GaN} = 90$ cm^2/Vs
$\mu_{AlN} = 30$ cm^2/Vs

Supplementary note: When discussing scattering mechanisms, we distinguish between ***momentum relaxation time*** and ***energy relaxation time***. In the formula $\mu = e\tau / m^*$, τ is the *momentum relaxation time*, as discussed earlier. What is the difference between the momentum and energy relaxation time?

Ionized impurity scattering is an *elastic* scattering mechanism and it can completely randomize ("relax") the momentum of an electron without affecting its energy. In contrast, phonon scattering is an *inelastic* scattering mechanism and thus leads to a momentum and energy loss of the electron. Therefore, the momentum-relaxation time is generally shorter than the energy-relaxation time. This consideration is relevant for velocity overshoot considerations in sub-micron FETs (see, for example, Chou *et al*., 1985).

19.4 Phenomenological mobility modeling

Experimental mobilities as a function of the doping concentration can be described with good accuracy by

$$\mu_n = \mu_{HC} + \frac{\mu_{LC} - \mu_{HC}}{1 + \left(N_D / N_{1/2}\right)^{2/3}} \tag{19.23}$$

where μ_{LC} is the low-concentration mobility, μ_{HC} is the high-concentration mobility, N_D is the n-type doping concentration, and $N_{1/2}$ is the concentration at which the mobility is reduced by approximately a factor of two as compared to the low-concentration mobility.

At low concentrations, the denominator in Eq. (19.23) has unit value and the mobility is given by μ_{LC}. At high concentrations, the denominator is approximately $(N_D / N_{1/2})^{2/3}$ so that the mobility decreases by a factor of $(10)^{2/3} \approx 4.6$ per order of magnitude of doping concentration. At very high concentrations, the mobility approaches μ_{HC}. Below, we give common fitting parameters for n- and p-type GaN:

n-type GaN:	$\mu_{LC} = 1800$ cm^2/Vs	$\mu_{HC} = 10$ cm^2/Vs	$N_{1/2} = 1.0 \times 10^{17}$ cm^{-3}
p-type GaN:	$\mu_{LC} \approx 40$ cm^2/Vs	$\mu_{HC} \approx 1$ cm^2/Vs	$N_{1/2} \approx 1.0 \times 10^{17}$ cm^{-3}

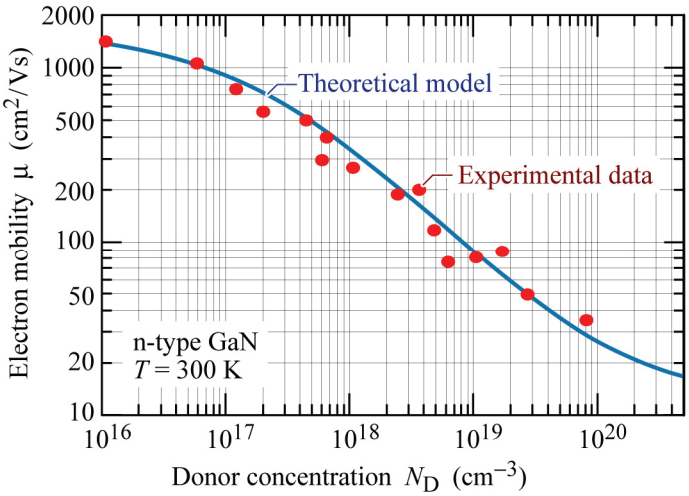


Fig. 19.11. Experimental electron mobilities measured at 300 K in n-type GaN and theoretical fit.

Theoretical model:

$\mu = \mu_{HC} + (\mu_{LC} - \mu_{HC}) / [1 + (N_D / N_{1/2})^{2/3}]$

where

$\mu_{LC} = 1800$ cm^2/Vs $\quad \mu_{HC} = 10$ cm^2/Vs

$N_{1/2} = 1 \times 10^{17}$ cm^{-3}

Figure 19.11 shows experimental electron mobilities in n-type GaN and the theoretical fit. Comparison of experiment and theory shows that the fitting formula works very well for n-type GaN at room temperature.

Additional mobility-versus-doping-concentration formulas, including the temperature dependence, can be found, for example, in device simulation software manuals (Silvaco, 1997). For the modeling of mobility, see also Piprek (2003).

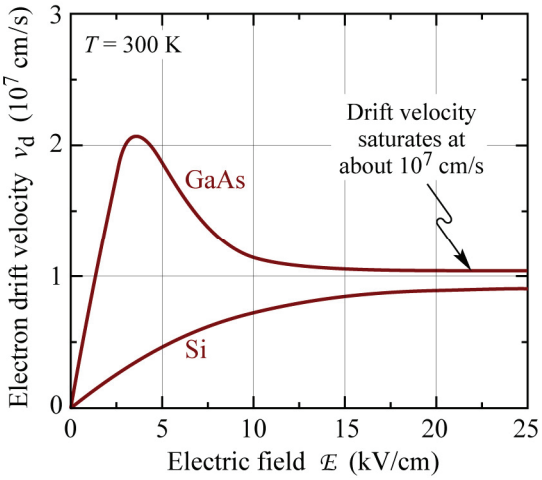


Fig. 19.12. Velocity-versus-electric-field characteristic of electrons in GaAs and Si at room temperature (after Sze, 1981).

19.5 Saturation velocity

While the drift velocity and the electric field are linearly related at *low* electric fields, the drift velocity saturates at *high* electric fields. This is shown for GaAs and Si in ***Fig***. 19.12. At fields of

$\mathcal{E} > 10$ kV/cm, the velocity is approximately 1×10^7 cm/s, and it does not increase further with the electric field.

We also note that the GaAs velocity-field characteristic has a maximum at the electric field of about 3 kV/cm and a regime of negative differential conductivity at about 5 kV/cm, before the velocity ultimately saturates.

The saturation of the drift velocity can be explained as follows: At very high fields, the carrier kinetic energy of rapidly propagating carriers becomes so large, that they are immediately scattered by optical phonons. At even higher fields, i.e. when carriers are more rapid accelerated, scattering by optical phonons occurs even more rapidly. This situation is shown in ***Fig***. 19.13. Inspection of the figure reveals that the drift velocity does not increase with electric field, i.e. *the drift velocity saturates*.

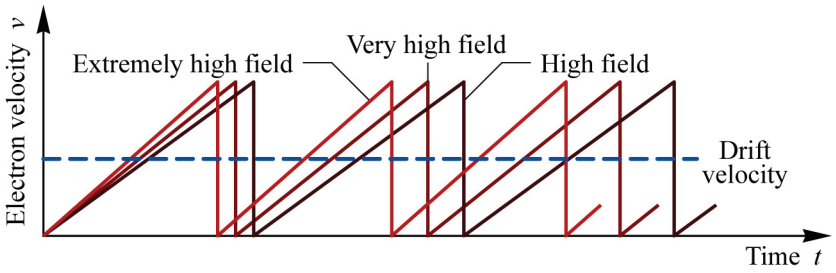


Fig. 19.13. Schematic velocity-verus-time characteristic in the high-field regime, where optical phonon scattering gives a saturated drift velocity, independent of the electric field.

Inspection of the GaAs velocity-field characteristic reveals that it has a maximum and a regime of negative differential resistance. This can be explained as follows: The conduction band structure of GaAs has, in addition to the central valley (the Γ valley at $k = 0$) an additional L and X valley, as illustrated in ***Fig***. 19.14 (a).

At high electric fields, carriers gain sufficient energy to populate the L valley, as illustrated in ***Fig***. 19.14 (b). The L valley has a lower curvature than the Γ valley. As a consequence, the electron mass in the L valley is much heavier and the mobility is much lower than in the Γ valley. In addition, the density of states in the L valley is much higher than in the Γ valley. These factors result in the L valley to become increasingly populated so that a negative differential resistance region occurs on the velocity-field characteristics of GaAs.

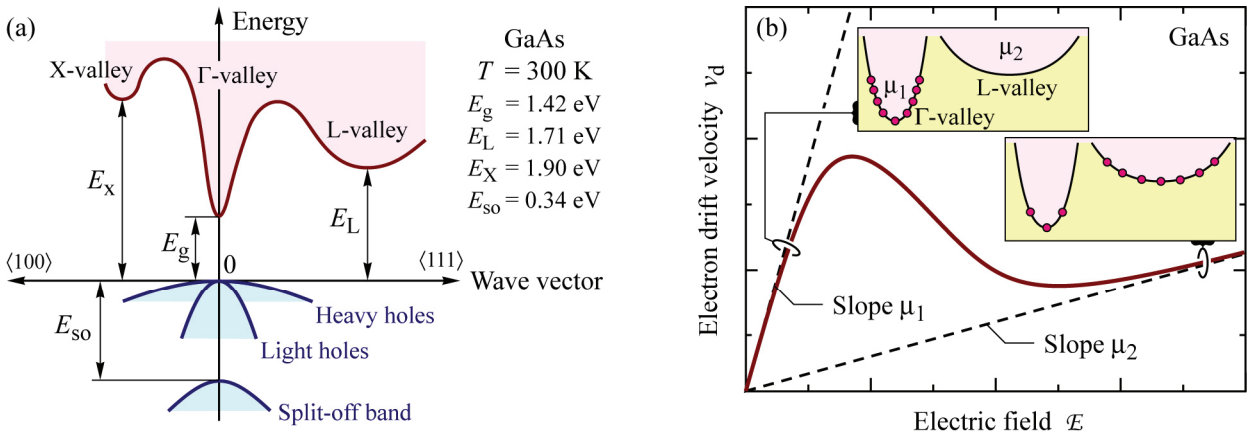


Fig. 19.14. (a) Band structure of GaAs. (b) Explanation of velocity-field relation of electrons in GaAs by population of high-mobility valley at low electric fields and by population of low-mobility satellite valley at high electric fields.

19.6 Diffusion (Brownian motion)

Whereas drift is a *directed* motion in an electric field, diffusion is the *random* movement of carriers that occurs due to their thermal energy. The random movement is caused by carriers having a kinetic energy of (3/2) kT. Only for low temperatures, $T \rightarrow 0$, the thermal motion comes to a halt.

Diffusion was first observed and reported by Robert Brown, a Scottish botanist in 1827 using the microscope shown in ***Fig***. 19.15 (a). Under the microscope minute particles, such as pollen or coal dust, apparently moved randomly around without cessation, as illustrated in ***Fig***. 19.15 (b). Brown was fascinated and, without understanding the movement, reported it to the scientific community. What was driving these particles to move about? Was there a secret force at work or was this, after all, the true origin of life? Other scientists reproduced the observation and confirmed the finding. Because no one knew what was driving the particles, the movement was called ***Brownian motion***. To date, we understand that Brown observed the diffusion of particles. However, at Brown's time, the physical process of diffusion was unknown, so that ***diffusion*** and ***Brownian motion*** are synonyms.

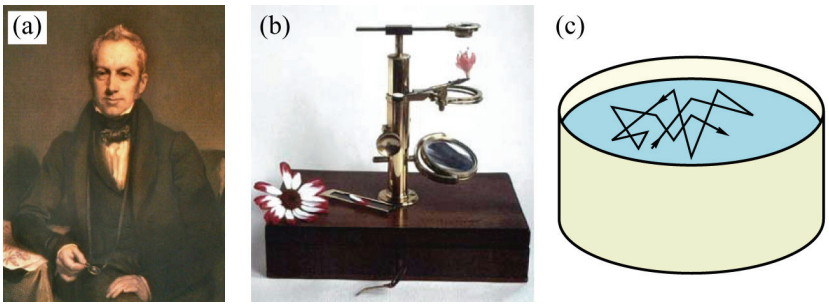


Fig. 19.15. (a) Robert Brown, Scottish botanist, 1773–1858 (after Australian National Botanic Gardens, 2006). (b) Microscope used by Brown (after Linnean Society, London, 2006). (c) Random movement of dust particles on a water surface.

Diffusion is a very basic phenomenon which occurs, for example, under the following circumstances:

- Diffusion of electrons and holes in semiconductors
- Diffusion of impurity atoms in semiconductors, especially at high temperatures
- Heat diffusing in a material
- Popular examples for diffusion are also the distribution of perfume scent in a room, a droplet if ink in a glass of water, or a droplet of milk in a cup of coffee.

The diffusion equation is given by:

$$\frac{\partial n}{\partial t} = D_n \nabla^2 n \tag{19.24}$$

where n is the electron concentration, D is the diffusion constant, and $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$. For a one-dimensional situation, the diffusion equation reduces to

$$\frac{\partial n(x)}{\partial t} = D_n \frac{\partial^2 n(x)}{\partial x^2} \,. \tag{19.25}$$

Let us assume that the carriers considered here, i.e. electrons, are minority carriers injected into a majority carrier region (p-type region). In this case electrons will disappear by recombining with holes. That is, electrons will change in concentration over time not only by diffusing away but also by recombining. In this case, an additional term is added to the diffusion equation representing the recombination of electrons. The diffusion equation is then given by

$$\frac{\partial n(x)}{\partial t} = D_n \frac{\partial^2 n(x)}{\partial x^2} - \frac{n(x)}{\tau_n} \tag{19.26}$$

where τ_n is the minority carrier lifetime.

Using the boundary condition $n(x<0) = \Delta n$ and $n(x \geq 0) = 0$, as shown in ***Fig***. 19.16, the solution of the diffusion equation is given by

$$\boxed{n(x) = \Delta n\, e^{-x/L_D}} \quad \text{with} \quad \boxed{L_D = \sqrt{D_n \tau_n}} \quad (\text{for } x \geq 0) \tag{19.27}$$

The solution is illustrated in ***Fig***. 19.16.

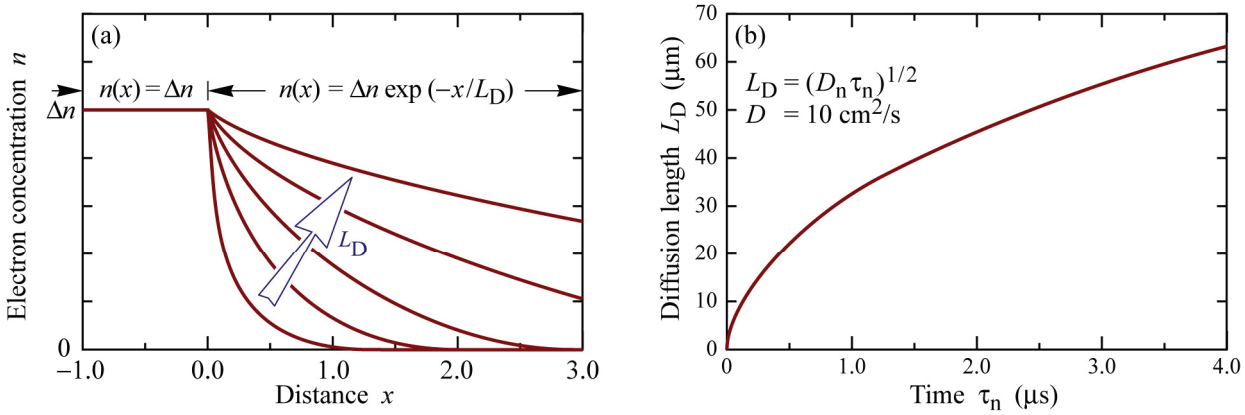


Fig. 19.16. (a) Diffusion of particles for different diffusion lengths. (b) Diffusion length for particles as a function of the minority carrier life time.

19.7 The Einstein relation

It is intuitively clear that a particle that diffuses rapidly will also drift rapidly under the influence of an electric field. This suggests that the diffusion constant and the mobility are proportional to each other. The diffusion constant of electrons and holes and their mobility are related by the Einstein relation which is give by

$$\boxed{D = \frac{kT}{e}\,\mu} \quad (\text{Einstein relation}) \tag{19.28}$$

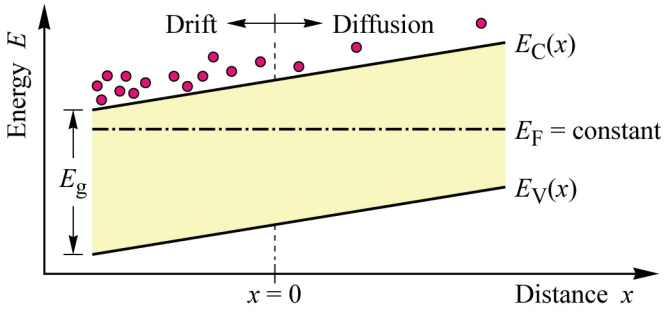


Fig. 19.17. Drift and diffusion of carriers in the depletion region of an unbiased pn junction.

We will prove the Einstein relation by considering electrons in an unbiased pn junction region where the carriers are subject to the internal electric field of the pn junction as shown in ***Fig***. 19.17. We will calculate the diffusion current and the drift current and use the condition that these currents must cancel under equilibrium conditions.

Diffusion current: Using the Boltzmann distribution for carriers, the carrier concentration is given by

$$n(x) \quad = \quad N_c \, e^{-(E_C(x)-E_F)/kT} \tag{19.29}$$

Because $E_C(x)$ is a function of position, a gradient of carriers exists, and the diffusion current density is given by

$$J_{\text{diffusion}} \quad = \quad e D_n \frac{dn(x)}{dx} \tag{19.30}$$

Insertion of Eq. (19.29) into Eq. (19.30) yields

$$J_{\text{diffusion}} \; = \; e D_n \frac{d}{dx} N_c \, e^{-(E_C(x)-E_F)/kT} \; = \; e D_n \, N_c \, e^{-(E_C(x)-E_F)/kT} \frac{-1}{kT} \frac{dE_C(x)}{dx} \tag{19.31}$$

Drift current: The drift current of electrons is given by

$$J_{\text{drift}} \quad = \quad e \, n(x) \, \mu_n \, \mathcal{E} \quad = \quad e \, N_c \, e^{-(E_C(x)-E_F)/kT} \, \mu_n \, \frac{1}{e} \, \frac{dE_C(x)}{dx} \tag{19.32}$$

where $\mathcal{E}$ is the electric field in the region considered here.

Using that under equilibrium conditions $|J_{\text{drift}}| = |J_{\text{diffusion}}|$, we can equate Eqs. (19.31) and (19.32) and obtain the Einstein relation $D = (kT/e)\,\mu$, which concludes our proof.

References

Adachi S. "GaAs, AlAs, and $Al_xGa_{1-x}As$: Material parameters for use in research and device applications" *Journal of Applied Physics* **58**, R1 (August 1985)

Brooks H. in *Advance in Electronics and Electron Physics*, edited by L. Marton (Academic Press, New York, NY, 1955) page 85 and156

Conwell E. and Weisskopf V. F. "Theory of Impurity Scattering in Semiconductors" *Physical Review* **77**, 388 (1950)

Chattopadhyay D. and Queisser H. J. "Electron scattering by ionized impurities in semiconductors" *Review of Modern Physics* **53**, 745 (1981)

Chou S. Y., Antoniadis D. A., and Smith H. I. "Observation of electron velocity overshoot in sub-100-nm-channel MOSFETs in silicon" *IEEE Electron Device Letters* **EDL-6**, 665 (1985)

Ieong, Meikei "Solid state devices and materials" http://www.ieong.net/ee3106 (2006)

Look D. C., Lorance D. K., Sizelove J. R., Stutz C. E., Evans K. R., and Whitson D. W. "Alloy scattering in p-type $Al_xGa_{1-x}As$" *Journal of Applied Physics* **71**, 260 (1992)

Matthiessen, Ludwig; An approximation known as Matthiessen's rule, was developed by the German physicist Ludwig Matthiessen (1830–1906). The rule asserts that all scattering mechanisms are independent (true only as long as the scattering mechanisms are isotropic and infrequent). Hence, the hindrances of each scattering mechanism can simply be added up. The electrical mobility in a semiconductor is then obtained from the the sum of all hindrances (scattering mechanisms). The year during which Matthiessen developed the rule is not known. (1906)

Morkoc H. *Nitride Semiconductors and Devices* (Springer Verlag, Berlin, Germany, 1999)

Piprek J. "Semiconductor optoelectronic devices: Introduction to physics and simulation" (Academic Press, San Diego, 2003)

Seeger K. *Semiconductor Physics* (Springer Verlag, Berlin, Germany, 1982)

Silvaco Corporation "Atlas User's Manual, Version 1.5.0" ATLAS Device Simulation Framework enables engineers to simulate the electrical, optical, and thermal behavior of semiconductor devices. ATLAS provides a physics-based, easy to use, modular, and extensible platform to analyze DC, AC, and time-domain responses for semiconductor based technologies in 2 and 3 dimensions. (1997)

Sze S. M. "Physics of semiconductor devices" 2nd edition (John Wiley and Sons, New York, 1981)

Wolfe C. M., Stillman G. E., and Lindley W. T. "Electron mobility in high-purity GaAs" *Journal of Applied Physics* **41**, 3088 (June 1970)

20

Selectively doped heterostructures

20.1 Selectively doped heterostructure basics

Selectively doped heterostructures, also called ***modulation-doped heterostructures***, are structures which consist of a doped wide-gap semiconductor and an undoped narrow-gap semiconductor. Selectively doped heterostructures were first realized by Stormer *et al*. (1978) and Dingle *et al.* (1978) in an attempt to reduce scattering of carriers by ionized impurities. The electron mobilities obtained in $Al_xGa_{1-x}As/GaAs$ heterostructures at low temperatures can exceed 10^7 cm^2/Vs (Pfeiffer *et al.*, 1989).

The band diagram of a selectively doped n-type heterostructure is shown in ***Fig***. 20.1 (a) and (b) before and after the electron transfer to the narrow-gap material, respectively. The structure consists of a doped wide-gap semiconductor and an undoped narrow-gap semiconductor. Because the conduction band edge of the narrow-gap semiconductor can be generally assumed to be lower in energy, electrons originating from donors in the wide-gap semiconductor transfer to the narrow-gap semiconductor. The transferred electrons form a quantized, two-dimensional electron gas (2DEG) located at the interface.

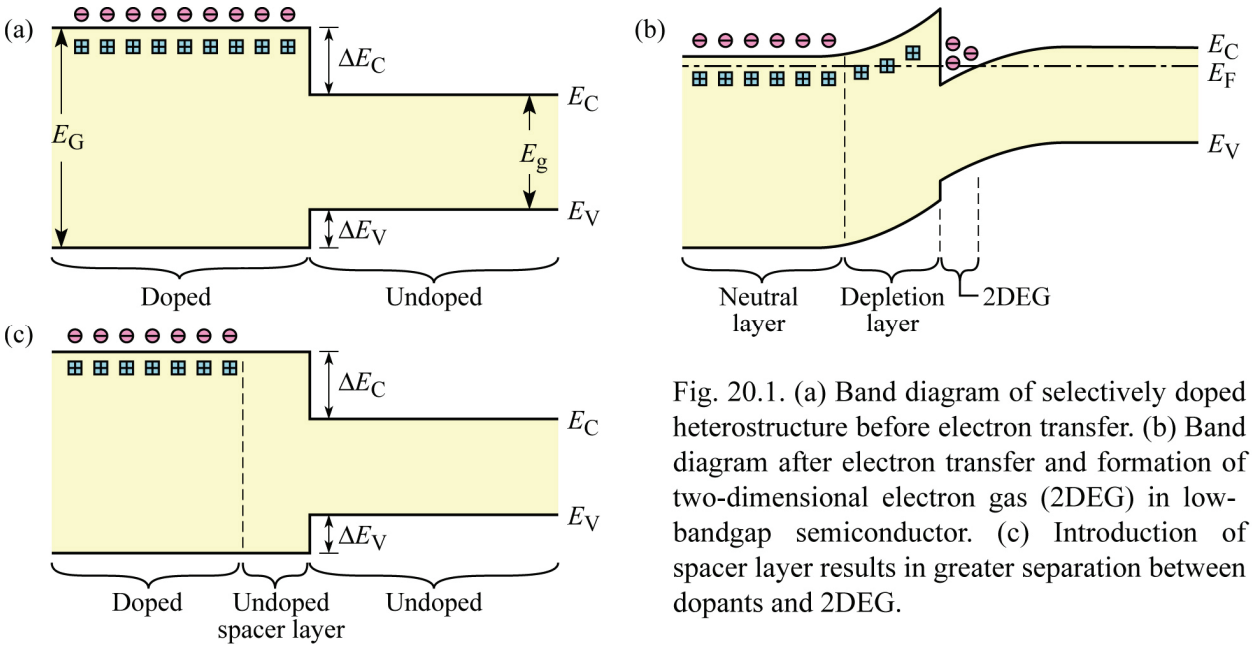


Fig. 20.1. (a) Band diagram of selectively doped heterostructure before electron transfer. (b) Band diagram after electron transfer and formation of two-dimensional electron gas (2DEG) in low-bandgap semiconductor. (c) Introduction of spacer layer results in greater separation between dopants and 2DEG.

Because electrons are spatially separated from their parent ionized impurities, ionized impurity scattering is reduced as compared to doped bulk semiconductors. The electron mobility is especially enhanced at low temperatures where ionized impurity scattering is the dominant scattering mechanism. At room temperature, the mobility enhancement is still significant thereby allowing semiconductor structures with otherwise unattainable high carrier mobilities.

The increase the electron-to-donor separation further, a spacer layer is introduced, as shown

in ***Fig***. 20.1 (c). Although the spacer layer increases the carrier mobility, it also reduces the electron concentration. Typical spacer layer thicknesses used in devices are 10 – 100 Å. A perspective view of the band diagram of a selectively doped heterostructure is shown in ***Fig***. 20.2.

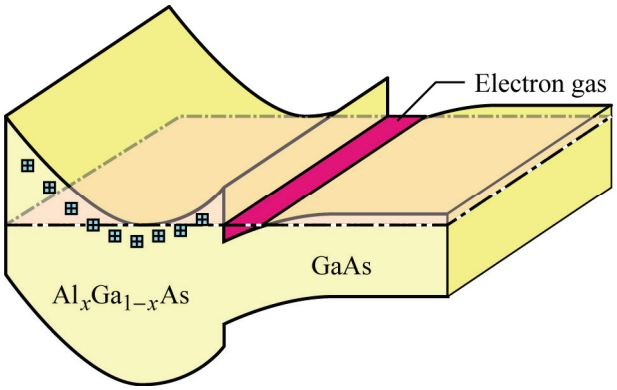


Fig. 20.2. Band diagram of selectively doped heterostructure. The two-dimensional electron gas (2DEG) forms in the low-bandgap semiconductor at the interface between the two semiconductors.

By spatially separating electrons from their parent ionized dopants, selectively doped heterostructures have reduces ionized impurity scattering. Because the ionized impurity mobility has a temperatures dependence of $T^{3/2}$, the mobility gain is particularly pronounced at low temperatures. This is evident from ***Fig***. 20.3 (a), which shows the schematic temperature dependence of the mobility of a bulk semiconductor and a selectively doped heterostructure. ***Figure*** 20.3 (b), shows the improvement that has been made during the period 1979 – 1989 when the pursuit of high electron mobilities, particularly at low temperatures, was a popular goal of many research groups. The increase in mobility in this period is testimony of advances in attaining lower background impurity concentrations and higher quality materials.

At room temperature, the mobility increase is not as pronounced as it is at low temperature. a At a doping concentration of 10^{17} cm^{-3} n-type GaAs has a typically mobility of 4000 cm^2/Vs. Selectively doped $Al_xGa_{1-x}As$/GaAs heterostructures have typical mobilities of 5000 – 8000 cm^2/Vs.

Exercise: ***Electron mobility in a selectively doped heterostructure***. Consider bulk GaAs with a phonon-scattering mobility of μ_{phonon} = 8000 cm^2/Vs. Assume an ionized-impurity-scattering mobility of μ_{II} = 8000 cm^2/Vs.

(a) What is the electron mobility in the bulk material?

(b) What would be the electron mobility in an equivalent selectively doped heterostructure?

Solution

(a) The measured mobility can be obtained from Matthiessen's rule according to $\mu^{-1} = \mu_{phonon}^{-1} + \mu_{II}^{-1}$ = (8000 cm^2/Vs) $^{-1}$ + (8000 cm^2/Vs) $^{-1}$ = (4000 cm^2/Vs) $^{-1}$. Thus the mobility of the bulk material is μ = 4000 cm^2/Vs.

(b) In a selectively doped heterostructure, ionized impurity scattering is absent and thus $\mu_{II} \to \infty$. Thus measured $\mu^{-1} = \mu_{phonon}^{-1} + \mu_{II}^{-1}$ = (8000 cm^2/Vs) $^{-1}$ + (∞ cm^2/Vs) $^{-1}$ = (8000 cm^2/Vs) $^{-1}$. Thus the mobility of the heterostructure would be μ = 8000 cm^2/Vs.

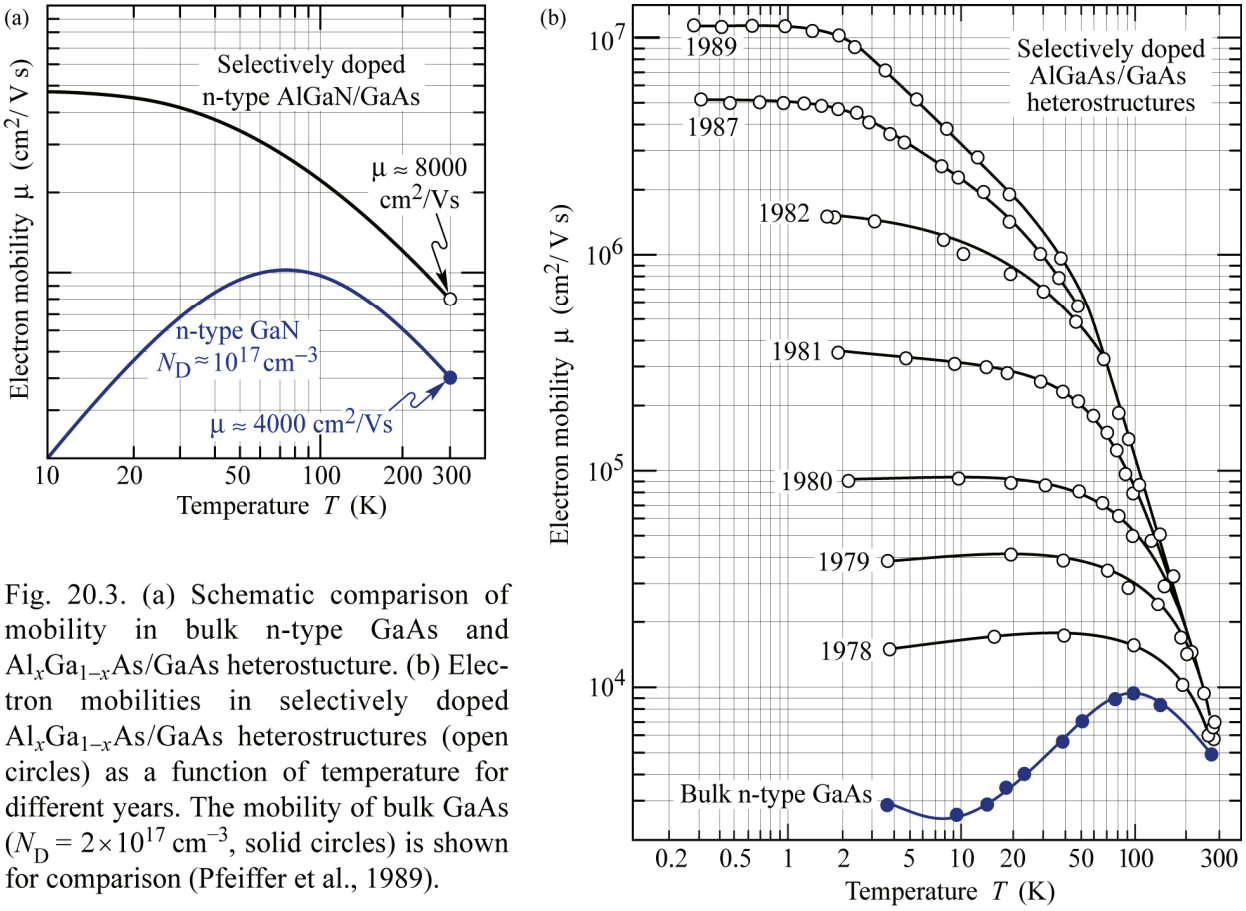


Fig. 20.3. (a) Schematic comparison of mobility in bulk n-type GaAs and $Al_xGa_{1-x}As$/GaAs heterostucture. (b) Electron mobilities in selectively doped $Al_xGa_{1-x}As$/GaAs heterostructures (open circles) as a function of temperature for different years. The mobility of bulk GaAs ($N_D = 2\times10^{17}$ cm^{-3}, solid circles) is shown for comparison (Pfeiffer et al., 1989).

20.2 Carrier concentration in selectively doped heterostructures

The two-dimensional (2D) electron density of the electron gas at the interface of the two semiconductors is determined by two driving forces. *First*, electrons transfer from their parent donor states to the narrow-gap semiconductor due to the availability of states at lower energy. *Second*, an electric dipole field is created by the electron transfer whose charges consist of the depleted donor layer in the wide-gap semiconductor and the electron gas in the narrow-gap material. The transfer of electrons continues until the Fermi level of the electron gas coincides with the donor energy in the neutral region of the wide-gap semiconductor.

The calculation of the electron density of a selectivity doped heterostructure is conveniently done by considering the energies of the electron states involved. The detailed conduction-band diagram of a selectively doped heterostructure is shown in ***Fig***. 20.4. As an example, we consider an $Al_xGa_{1-x}As$/GaAs heterostructure consisting of a doped $Al_xGa_{1-x}As$ region, an undoped $Al_xGa_{1-x}As$ spacer, and an undoped GaAs narrow-gap region. The purpose of the undoped $Al_xGa_{1-x}As$ spacer is to further spatially separate the free electrons from their parent ionized impurities. This separation was shown to further increase the electron mobility (Stormer *et al.*, 1981). Next we consider the different energies involved in the heterostructure.

(i) Donor ionization energy, E_d: The thermal ionization energy of the shallow donor in GaAs and $Al_xGa_{1-x}As$ is approximately 5 meV. At low temperatures, the Fermi level coincides with the donor level in the un-depleted $Al_xGa_{1-x}As$ side of the heterostructure.

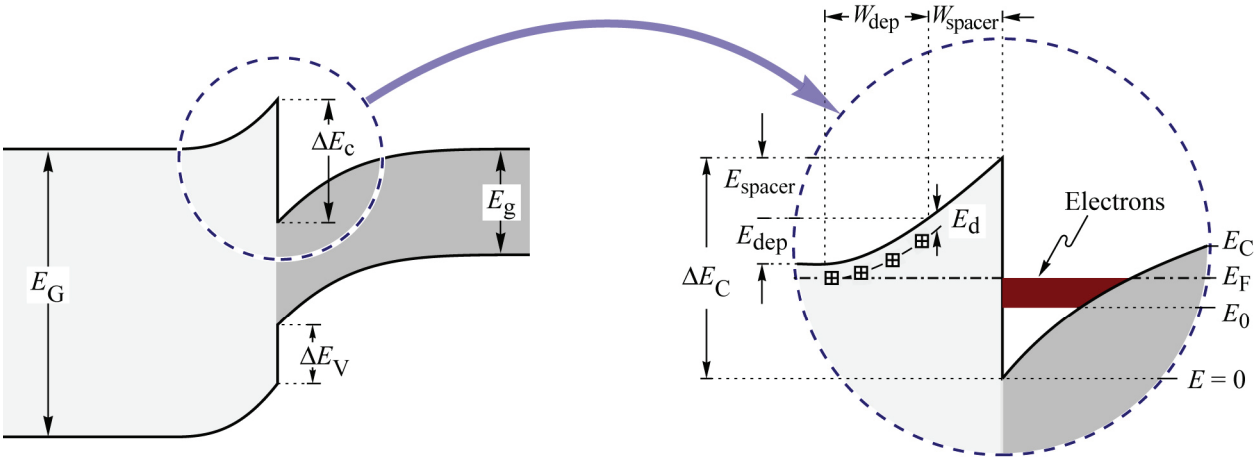


Fig. 20.4. Band diagram of a selectively doped heterostructure consisting of a doped wide-gap semiconductor and an undoped narrow-gap semiconductor. Due to the lower conduction band energy, electrons transfer from their parent donors to the narrow-gap semiconductor.

(ii) Depletion energy, E_{dep}: Assuming a three-dimensional (3D) doping concentration N_D, the 2D density of transferred electrons is given by

$$n_{2DEG} = N_D W_{dep} \tag{20.1}$$

where W_{dep} is the (unknown) width of the depletion region. Any residual acceptor concentration is neglected. The magnitude of the electric field at the end of the depletion region can be obtained from Gauss's law

$$\mathcal{E} = \frac{e}{\varepsilon} N_D W_{dep} = \frac{e}{\varepsilon} n_{2DEG} \tag{20.2}$$

The energy drop in the depletion region is then given by

$$E_{dep} = \frac{1}{2} e \mathcal{E} W_{dep} = \frac{e^2}{2\varepsilon} \frac{n_{2DEG}^2}{N_D} \tag{20.3}$$

In the case of $N_D \to \infty$, *i. e.* for δ-doping, the depletion region thickness and energy drop approach zero, *i. e.* $W_{dep} \to 0$ and $E_{dep} \to 0$.

(iii) Spacer energy, E_{spacer}: The energy drop in the spacer region is given by

$$E_{spacer} = e \mathcal{E} W_{spacer} = \frac{e^2}{\varepsilon} n_{2DEG} W_{spacer} \tag{20.4}$$

where W_{spacer} is the width of the spacer region.

(iv) Conduction band discontinuity, ΔE_c: The conduction band discontinuity is a materials parameter. For the material system $Al_xGa_{1-x}As$/GaAs the conduction band discontinuity is given by

$$\Delta Ec \approx \frac{70}{100}\Delta E_g = \frac{70}{100}1.247\,\text{eV}\ x_{Al} \tag{20.5}$$

where it is assumed that 70% of the bandgap discontinuity occurs in the conduction band. The factor 1.247 eV x_{Al} expresses the change of the $Al_xGa_{1-x}As$ band gap with Al mole fraction (Casey and Panish, 1978). For a 32% Al mole fraction, the conduction band discontinuity calculated from Eq. (20.5) equals 279 meV.

(v) Subband energy, E_0: The ground-state energy of the triangular well can be obtained by a variational calculation using the trial function

$$\psi(z) = A\,z\,e^{-\alpha z} \tag{20.6}$$

where α is the trial parameter and A is a normalization parameter. The wave function vanishes at $z = 0$, *i. e.* at the interface, as schematically shown in ***Fig***. 20.5. The wave function decays exponentially in the triangular GaAs side of the barrier. The normalization condition $\langle \Psi | \Psi \rangle = 1$ yields the normalization constant $A = 2\,\alpha^{3/2}$ and thus the normalized wave function

$$\psi(z) = 2\alpha^{3/2}\,z\,e^{-\alpha z} \tag{20.7}$$

which is also called the Fang-Howard wave function (Fang and Howard, 1966).

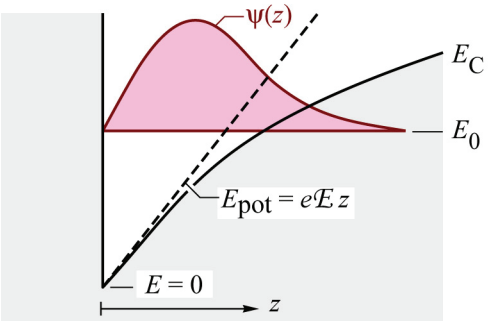


Fig. 20.5. Fang–Howard wave function, $\psi(z) \propto z\,\exp(-\alpha z)$, at the interface of an $Al_xGa_{1-x}As$/GaAs heterostructure.

Calculation of the energy expectation value for this wave function yields

$$\langle E \rangle = \langle \psi | H | \psi \rangle = \left\langle \psi \left| -\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2} + E_{pot}(z) \right| \psi \right\rangle = e\,\mathcal{E}\,\frac{3}{2\alpha} + \frac{\hbar^2}{2m^*}\alpha^2 \tag{20.8}$$

where the potential energy $E_{pot}(z) = e\,\mathcal{E}z$. Minimizing the expectation value of the energy yields the trial parameter α according to

$$\alpha = \left[\frac{3}{2}\,e\,\mathcal{E}\,\frac{m^*}{\hbar^2}\right]^{1/3} \tag{20.9}$$

Insertion of the trial parameter α into Eq. (20.8) yields the ground-state energy of the triangular potential well

$$E_0 \quad = \quad \langle E \rangle_{\min} \quad = \quad \frac{3}{2}\left[\frac{3}{2}\, e\, \mathcal{E}\, \frac{\hbar}{\sqrt{m^*}}\right]^{2/3} \tag{20.10}$$

which can also be expressed in terms of the 2D electron density

$$E_0 \quad = \quad \frac{3}{2}\left(\frac{3}{2}\frac{e^2}{\varepsilon}\, n_{2\mathrm{DEG}}\, \frac{\hbar}{\sqrt{m^*}}\right)^{2/3} \tag{20.11}$$

(vi) Degeneracy energy of 2DEG, $E_F - E_0$. Due to the finite density of states, the Fermi level increases above the ground-state energy E_0 at high free carrier densities. The energy $E_F - E_0$ is obtained as

$$E_F - E_0 \quad = \quad n_{2\mathrm{DEG}} / \rho_{\mathrm{DOS}}^{2\mathrm{D}} \tag{20.12}$$

where ρ_{DOS}^{2D} is the two-dimensional density of states given by $\rho_{DOS}^{2D} = m^*/(\pi\,\hbar^2)$. The equation is valid for degenerate electron systems, where the Fermi – Dirac distribution can be approximated by a step function.

The total energy balance of the heterostructure can be written as (see ***Fig***. 20.4 for illustration)

$$\boxed{E_d \;+\; E_{dep} \;+\; E_{spacer} \;+\; E_0 \;+\; \left(E_F - E_0\right) \quad = \quad \Delta E_c} \tag{20.13}$$

where it is assumed that the Fermi energy is the same on both sides of the interface, *i. e.* the system is in equilibrium. All the energies in Eq. (20.13) can be expressed as a function of n_{2DEG} which is then the only variable in the equation. It is not possible to explicitly solve the equation for n_{2DEG}. However, the solution for n_{2DEG} can be easily obtained numerically.

Fully self-consistent calculations of the electron density in selectively doped heterostructures have been reported in the literature (Ando, 1982a and 1982b; Stern and Das Sarma, 1984; Vinter, 1984). In such self-consistent calculations, the Schrödinger and Poisson equations are solved in an iterative process and many-body effects may be taken into account. Nevertheless, the results obtained from the approximate calculation described above compare favorably with the much more elaborate self-consistent calculations. In self-consistent calculations, the solution of Poisson's equation is straightforward whereas the accurate solution of Schrödinger's equation is more elaborate. Therefore, several groups used approximate solutions of the Schrödinger equation, including the variational solution (Stern, 1983; Fang and Howard, 1966) and the WKB approximation (Ando, 1985).

As an example, the carrier density of the 2DEG calculated from Eq. (20.13) is shown in ***Fig***. 20.6 as a function of the spacer thickness. The Al mole fraction used is $x = 32\%$ which leads to a conduction band discontinuity of $\Delta E_C = 279$ meV. Other parameters used are $\varepsilon/\varepsilon_0 = \varepsilon_r = 13.1$ and $m^* = 0.067 m_0$. The structure is assumed to be doped at different doping levels as well as δ-doped which results in a zero depletion energy $E_{dep} = 0$. ***Figure*** 20.6 reveals that the electron density decreases for large spacer thicknesses. The values of n_{2DEG} are lower for homogeneous doping as compared to the δ-doped case, as shown in ***Fig***. 20.6. Delta-doped heterostructures were shown to have higher electron densities (Schubert *et al.*, 1987) as well as higher electron mobilities (Schubert *et al.*, 1989) as compared to their homogeneously doped counterparts.

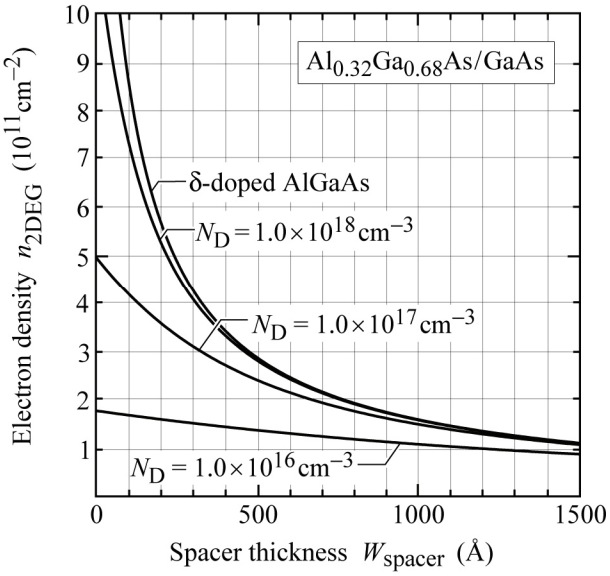


Fig. 20.6. Calculated electron density in selectively doped $Al_xGa_{1-x}As/GaAs$ heterostructures as a function of spacer thickness for different doping concentrations.

$Al_xGa_{1-x}As/GaAs$
$\Delta E_g = 1.247\ x$ eV
$\Delta E_C = 0.70\ \Delta E_g$
$\varepsilon_{r,AlGaAs} = 13.1 - 3.0\ x$
$E_d \approx 0$ meV (shallow donor)

20.3 Parallel conduction in selectively doped heterostructures

Whereas the 2DEG mobility is high, it is low for the doped n-type $Al_xGa_{1-x}As$ layer. Therefore, to attain high mobilities, it must be ensured that the doped $Al_xGa_{1-x}As$ layer is depleted of free electrons and thus does not participate in the electron transport. ***Figure*** 20.7 (a) shows the band diagram of a selectively doped heterostructure in which the $Al_xGa_{1-x}As$ layer is completely depleted. In contrast, ***Fig***. 20.7 (b) shows a heterostructure in which the $Al_xGa_{1-x}As$ layer is not completely depleted so that a second (parallel) conduction channel is formed in the $Al_xGa_{1-x}As$.

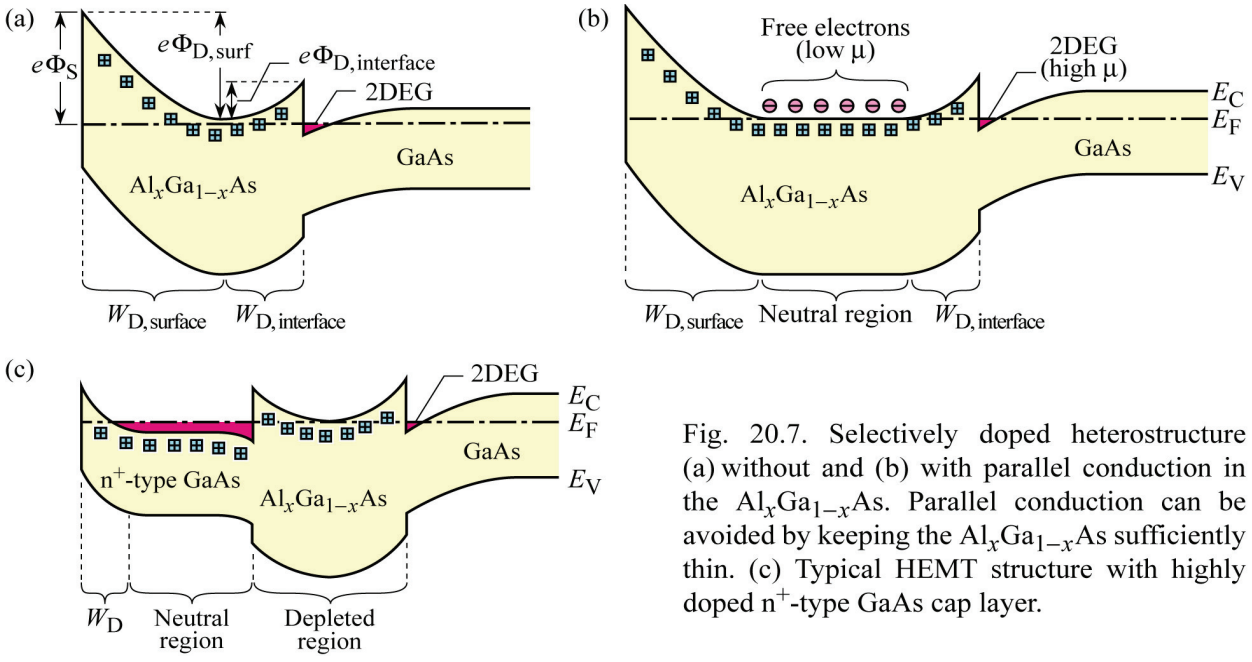


Fig. 20.7. Selectively doped heterostructure (a) without and (b) with parallel conduction in the $Al_xGa_{1-x}As$. Parallel conduction can be avoided by keeping the $Al_xGa_{1-x}As$ sufficiently thin. (c) Typical HEMT structure with highly doped n^+-type GaAs cap layer.

The figure illustrates the importance of having exactly the right thickness of the $Al_xGa_{1-x}As$ layer. That is, the $Al_xGa_{1-x}As$ thickness should be equal to the sum of the surface-depletion layer and interface depletion layer thickness. If the $Al_xGa_{1-x}As$ layer is too thick, parallel conduction

occurs. If the layer is too thin, the electron concentration in the 2DEG decreases. In both cases, the performance of field-effect transistors, made from the heterostructure, suffers.

Figure 20.7 (c) shows the band diagram of a selectively doped heterostructure in which a heavily doped n^+-type GaAs cap layer has been included. Such heavily doped cap layers are used to reduce parasitic resistances in high-electron-mobility transistors (HEMTs) with a recessed gate structure, as will be further discussed in the following section.

20.4 High electron mobility transistors

High electron mobility transistors (***HEMTs***), which are also known by the name of ***modulation-doped transistors*** (***MODFETs***) and ***heterostructure field-effect transistors*** (***HFETs***), were, when introduced in the 1980s, a new generation of transistors. They are based on high-mobility selectively doped heterostructures. Due to the clear advantages of HEMTs, they revolutionized compound semiconductor field-effect transistors. To date, most high-performance compound semiconductor field-effect transistors are based on the HEMT principle.

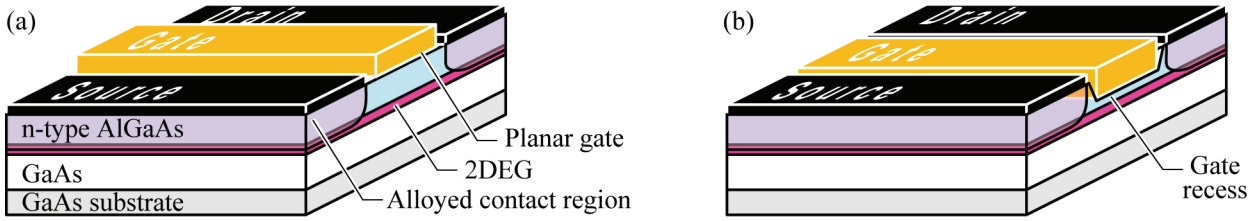


Fig. 20.8. (a) Basic structure of the high-electron mobility transistor (HEMT) with (a) planar and (b) recessed-gate structure.

The HEMT structure consists of the conventional source, gate and drain electrodes, as shown in ***Fig***. 20.8. The conduction between the source and the drain is provided by the two-dimensional electron gas (2DEG), i.e. the electron channel. The gate structure can be either *planar* or *recessed* as shown in ***Fig***. 20.8 (a) and ***Fig***. 20.8 (b), respectively. The recessed gate structure has the advantage of a lower access resistance but the disadvantage of an additional processing step, i.e. the gate-recess etch step.

Due to the high mobility of electrons in the channel, the HEMT is a fast transistor. How does the high electron mobility make the HEMT a fast transistor? To answer this question, we recall that there are two basic models for the operation of an FET, one being ***Shockley's gradual-channel-approximation model*** and the other one being the ***saturated-velocity model***.

Without mathematical derivation, we give the transconductance of a MOSFET-like structure based on Shockley's gradual-channel-approximation model:

$$g_{\mathrm{m,\,sat}} = \frac{\mathrm{d}I_{\mathrm{D,\,sat}}}{\mathrm{d}V_{\mathrm{GS}}} = \frac{\varepsilon_{\mathrm{WG}}\,\mu\,Z}{d_{\mathrm{WG}}\,L_{\mathrm{G}}}\left(V_{\mathrm{GS}} - V_{\mathrm{th}}\right) \qquad \text{(Shockley model)} \qquad (20.14)$$

where $\varepsilon_{\mathrm{WG}}$ and d_{WG} is the electrical permittivity and thickness of the wide-gap material, respectively, and the other symbols have their usual meaning. The Shockley model makes clear what is needed for a high-transconductance transistor, namely

- A high electron mobility, μ
- A small distance between the gate electrode and the electron channel, d_{WG}

Both characteristics are afforded by the HEMT.

The realization that the electric field under the gate electrode is very high has motivated the saturated velocity model, which assumes that electrons drift with the saturated velocity model under the gate electrode. Without mathematical derivation, we give the transconductance of a MOSFET-like structure based on the saturated velocity model:

$$g_{\mathrm{m,\,sat}} = \frac{\mathrm{d}I_{\mathrm{D,\,sat}}}{\mathrm{d}V_{\mathrm{GS}}} = -\frac{\varepsilon_{\mathrm{WG}}}{d_{\mathrm{WG}}}\, v_{\mathrm{sat}}\, Z \qquad \text{(Saturated velocity model)} \qquad (20.15)$$

This model indicates that the transconductance of a HEMT depends only on the saturated velocity and *not* on the mobility! It is important to keep in mind that both theoretical models are idealized, one assuming that the electron drift is purely mobility controlled, whereas the other one assuming that the electron drift is purely saturation-velocity controlled. Certainly the Shockley model has its limitations: The increase in electron mobility by, say a factor of 10, which can be achieved at low temperatures, does not result in an increase of g_m by a factor of 10. Certainly the saturated-velocity model also has its limitations in that experiments generally show that a higher mobility does indeed increase the g_m of a transistor. Therefore we can state that g_m benefits from the HEMT's high mobility, but we would not go as far as stating that g_m and μ are directly proportional.

It is instructive to also examine the parasitic resistive elements of a HEMT. ***Figure*** 20.9 distinguishes between the ***intrinsic*** or ***inner FET*** and the ***extrinsic*** or ***outer FET*** that can be accessed by external electrodes. The HEMT has a clear advantage in terms of a lower access resistance. Particularly the resistor R_{ch} is very low for the HEMT because of the high channel mobility. Note that the electric field is much lower outside the inner FET, thereby taking full advantage of the high mobility.

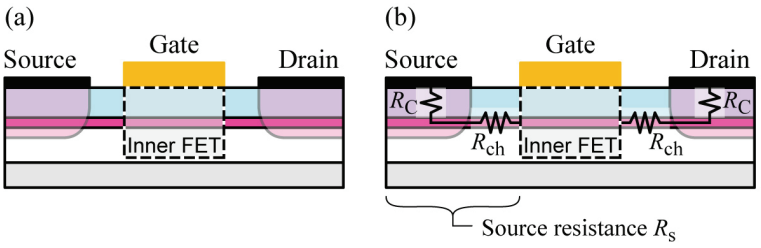


Fig. 20.9. (a) Basic structure and (b) resistive elements of the HEMT. R_{ch} and R_{C} are "access resistances" because they provide the access to the inner FET.

The low access resistance of the HEMT has two consequences: *Firstly*, low parasitic access resistances (particularly a low source resistance) result in a high transconductance. *Secondly*, low parasitic access resistances increase the device power-efficiency. *Thirdly*, every resistor creates white noise, so that the HEMT with its low access resistance exhibits excellent low-noise performance.

The above discussion allows us to establish the following design principles for a HEMT:

- Attain high concentration of the 2DEG by keeping spacer thickness small and by using a high doping concentration in the wide-gap material.
- Attain high mobility by a large spacer-layer thickness (typical spacer-layer thicknesses of HEMTs are 10 – 50 Å).
- Attain low access resistances by using a gate recess. The access resistance connects the outer FET accessible by probes with the inner FET.
- Use doped wide-gap semiconductor with high energy gap (i.e. increase Al content, x, in $Al_xGa_{1-x}As$/GaAs or $Al_xGa_{1-x}N$/GaN) to increase ΔE_C and thus the 2DEG density.

- Use undoped narrow-gap semiconductor with low energy gap (i.e. increase In content, y, in $Al_xGa_{1-x}As/Ga_{1-y}In_yAs$ or $Al_xGa_{1-x}N/Ga_{1-y}In_yN$) to increase ΔE_C and thus the 2DEG density.
- Use a binary compound for spacer layer (e.g. AlAs or AlN) to reduce alloy scattering.
- Use a *quantum well* heterostructure rather than a *single-interface* heterostructure and dope the top and bottom barrier layer of the quantum well. This results in a higher 2DEG density than the single-interface heterostructure that is doped only from one side.

Exercise: ***Reduction of access resistance in HEMT by recessed gate structure***. Explain how the access resistance is reduced in a HEMT that employs a recessed gate structure.

Solution: In the non-recessed region, parallel conduction occurs so that electrons can flow to the inner FET through the heavily doped n^+-type GaAs cap layer as well as through the 2DEG, as shown in ***Fig***. 20.10. In the gate-recess region, the n^+-type GaAs cap layer is etched away and the $Al_xGa_{1-x}As$ is completely depleted leaving only the high-mobility electron channel, as shown in ***Fig***. 20.10.

Alternatively, a high conductivity in the non-recessed region can also be created by ion implantation into the non-recessed region.

To attain a good etch-depth control in the recessed gate region, selective wet etching in combination with an AlAs etch-stop layer is frequently employed.

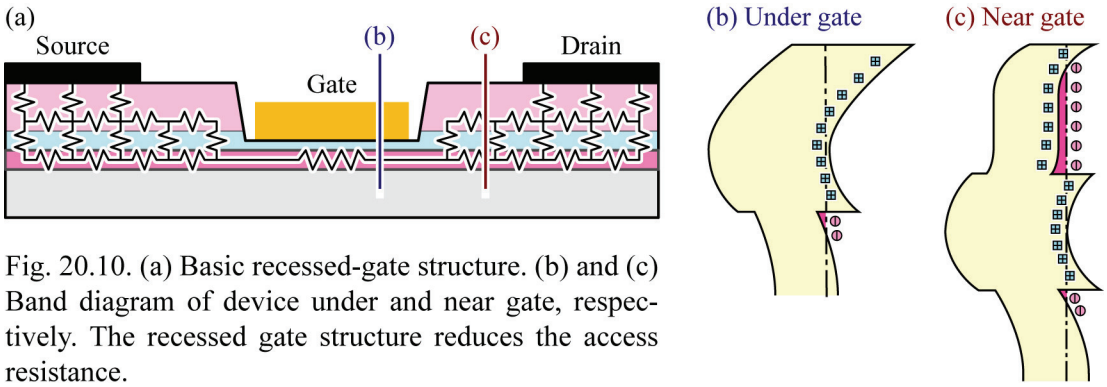


Fig. 20.10. (a) Basic recessed-gate structure. (b) and (c) Band diagram of device under and near gate, respectively. The recessed gate structure reduces the access resistance.

Exercise: ***Gain compression***. The output characteristic of a HEMT is shown in ***Fig***. 20.11. Inspection of the figure reveals that the gain (transconductance) is lower at V_{DS} = 20 V than it is at V_{DS} = 8 V. This phenomenon is called ***gain compression***. What may be the reason for gain compression?

Solution: One possible explanation for gain compression is device heating, which reduces the electron mobility and thus the gain. Another possible explanation of gain compression is velocity saturation at high electric fields under the gate.

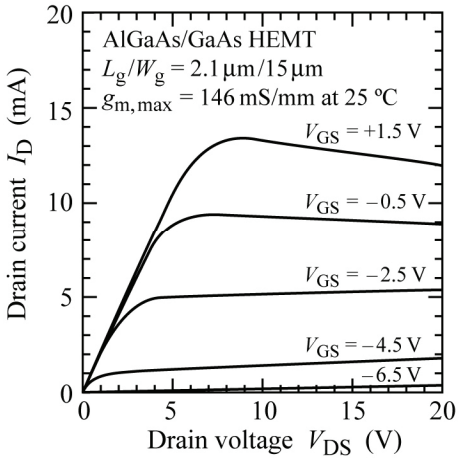


Fig. 20.11. Output characteristic of an AlGaAs/GaAs HEMT showing "gain compression" at high drain voltages (after Nogoya Institute of Technology, Japan, 2002).

Exercise: ***The mushroom gate***. A gate with a T-shaped or mushroom-shaped gate cross section is shown in ***Fig***. 20.12. What is the advantage of the mushroom gate?

Solution: A mushroom-shaped gate has a much larger cross section than a conventional gate electrode thereby reducing the resistance of the gate metal. A mushroom shaped gate is particularly desirable for devices with a very short (sub-µm) gate lengths, which would otherwise be very resistive.

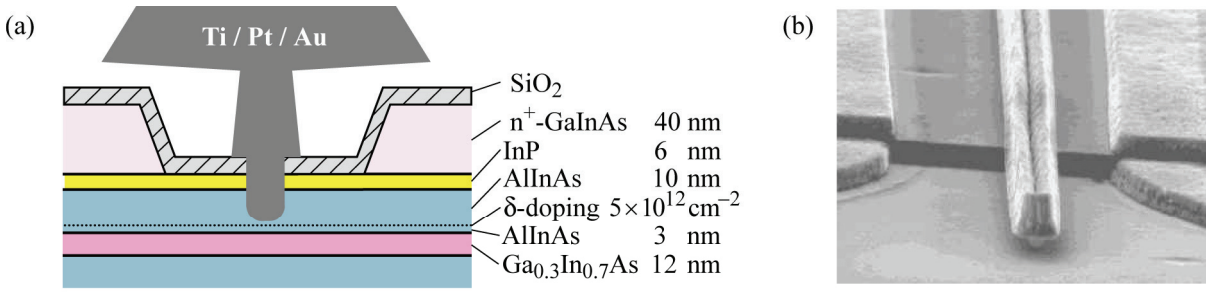


Fig. 20.12. (a) Cross section of pseudomorphic AlInAs/GaInAs HEMT with T-shaped gate electrode, also called "mushroom gate" (after Fujitsu Corporation, Japan, 2002). (b) Mushroom gate (after Nanyang Technological University, Singapore, 2006).

References

Ando T. "Self-consistent results for a GaAs/$Al_xGa_{1-x}As$ heterojunction. I. Subband structure and light-scattering spectra" *Journal of the Physical Society of Japan* **51**, 3893 (1982a)

Ando T. "Self-consistent results for a GaAs/$Al_xGa_{1-x}As$ heterojunction. II. Low temperature mobility" *Journal of the Physical Society of Japan* **51**, 3900 (1982b)

Ando T. "Subbands in Space-Charge Layers on Narrow Gap Semiconductors: Validity of Semiclassical Approximation" *Journal of the Physical Society of Japan* **54**, 2676 (1985)

Casey H. C. and Panish M. B. in *Heterostructure Lasers Part B - Materials and operating characteristics*, p.16 (Academic Press, New York, 1978)

Dingle R., Stormer H. L., Gossard A. C., and Wiegmann W. "Electron mobilities in modulation-doped semiconductor heterojunction superlattices" *Applied Physics Letters* **33**, 665 (1978)

Fang F. F., and Howard W. E. "Negative Field-Effect Mobility on (100) Si Surfaces" *Physical Review Letters* **16**, 797 (1966)

Pfeiffer L., West K. W., Stormer H. L., and Baldwin K. W. "Electron mobilities exceeding 10^7 cm^2/V s in modulation-doped GaAs" *Applied Physics Letters* **55**, 1888 (1989)

Schubert E. F., Cunningham J. E., Tsang W. T., and Timp G. L. "Selectively δ-doped $Al_xGa_{1-x}As$/GaAs heterostructures with high two-dimensional electron-gas concentrations $n_{2DEG}1.5\times10^{12}$ cm^{-2} for field-effect transistors" *Applied Physics Letters* **51**, 1170 (1987)

Schubert E. F., Pfeiffer L., West K. W., and Izabelle A. "Dopant distribution for maximum carrier mobility in selectively doped $Al_{0.30}Ga_{0.70}As$/GaAs heterostructures" *Applied Physics Letters* **54**, 1350 (1989)

Stormer H. L., Dingle R., Gossard A. C., Wiegman W., and Logan R. A. "Electronic properties of modulation doped GaAs-$Al_xGa_{1-x}As$" *Institute of Physics Conference Series* **43**, p. 557, edited by B. L. H. Wilson (Institute of Physics, London, 1978)

Stormer H. L., Pinczuk A., Gossard A. C., and Wiegmann W. "Influence of an undoped (AlGa)As spacer on mobility enhancement in GaAs-(AlGa)As superlattices" *Applied Physics Letters* **38**, 691 (1981)

Stern F. "Doping considerations for heterojunctions" *Applied Physics Letters* **43**, 974 (1983)

Stern F. and Das Sarma S. "Electron energy levels in GaAs-$Ga_{1-x}Al_xAs$ heterojunctions" *Physical Review* **B30**, 840 (1984)

Vinter B. "Subbands and charge control in a two-dimensional electron gas field-effect transistor" *Applied Physics Letters* **44**, 307 (1984)