Physics of Semiconductors and Nanostructures

Debdeep Jena (djena@cornell.edu)
Cornell University

April 16, 2017

Contents

1 And off we go! 7

1.1 Beyond belief 7
1.2 A brief history of Semiconductors 7
1.3 Future 8
1.4 These boots are made for walking 9
Summary 9
Problems 9

2 Secrets of the classical electron 11

2.1 Our ancestors knew metals 11
2.2 Discovery of the electron and its aftermath 11
2.3 Drude's model explains Ohm's law 12
2.4 Metals are Shiny 13
2.5 Metals conduct heat 13
2.6 Icing on the cake: The Weidemann-Franz Law 14
2.7 All is *not* well 14
Summary 15
Problems 16

3 Quantum Mechanics in a Nutshell 17

3.1 Photons 17
3.2 Wave-Particle Duality 18
3.3 The wavefunction 19
3.4 Operators 20
3.5 States of definite momentum and location 21
3.6 States of definite energy: The Schrodinger equation 23
3.7 Time-dependent Schrodinger equation 23
3.8 Stationary states and time evolution 25
3.9 Quantum Current 25
3.10 Fermions and Bosons 27
3.11 Spin, and the Spin-Statistics Theorem 29
3.12 The Dirac equation and the birth of particles 29
Summary 30
Problems 30

4 Damned Lies, and Statistics 31

4.1 The physics of equilibrium 31
4.2 Partition Function for Quantum Systems 32
4.3 The Fermi-Dirac Distribution 33
4.4 The Bose-Einstein Distribution 34
4.5 Manifestations of the distribution functions 34
4.6 Meaning of equilibrium in semiconductor devices 37
Summary 40
Problems 41

5 Electrons in the quantum world 43
5.1 In Schrodinger's equation we trust 43
5.2 The free electron 44
5.3 Not so free: particle on a ring 45
5.4 The electron steps into a higher dimension: 2D 48
5.5 Electrons in a 3D box 49
5.6 Resolving Drude's dilemma in the Quantum World 51
5.7 The particle in a box 51
5.8 The Dirac-Delta Function 51
5.9 The harmonic oscillator 52
5.10 The Hydrogen atom 52
5.11 Origin of the Elements: The Periodic Table 52
5.12 Origin of the Chemical Bond 53
5.13 Electrons in a periodic potential: Bloch Theorem 53

6 Red or Blue pill: Befriending the Matrix 55
6.1 The Expansion Principle 55
6.2 Matrix Mechanics 56
6.3 Matrices and Algebraic Functions 58
6.4 Looking ahead 61
Summary 61
Problems 62

7 Perturbations to the Electron's Freedom 63
7.1 Degenerate Perturbation Theory 63
7.2 Non-Degenerate Perturbation Theory 65
7.3 The Brillouin-Wigner Perturbation Results 68
7.4 Rayleigh-Schrodinger Perturbation Results 68
7.5 The Hellmann Feynman Theorem 70
7.6 Perturbation Theory Example 70
Summary 72
Problems 72

8 Electrons in a Crystal get their Bands, Gaps and Masses 75
8.1 The free-electron 76
8.2 Periodic perturbation 76
8.3 Degenerate Perturbation Theory 77
8.4 Non-degenerate Perturbation Theory 81
8.5 Glimpses of the Bloch Theorem 82
8.6 Non-periodic potentials and scattering 82
Summary 83
Problems 84

9 The k dot p Bandstructure 85
9.1 Perturbation theory 85
9.2 Symmetry 86
9.3 $\mathbf{k} \cdot \mathbf{p}$ theory 88
9.4 No spin-orbit interaction 88
9.5 With spin-orbit interaction 91
9.6 Further reading 93

10 Exact Bandstructure and Going Green 95
10.1 Exact Solution: The Kronig-Penney Model 95
10.2 Tight-binding models emerge from Kronig-Penney 99
10.3 Point defects in Kronig-Penney Models 101
10.4 Green's functions and Kronig-Penney for higher-dimensions . 103
10.5 Branch points of bandstructure and band-alignments 105

11 Doping and Semiconductor Heterostructures: The Effective Mass Method 107
11.1 Effective Mass Approximation, Envelope Functions 107
11.2 3D, 2D, 2D, 0D: Heterostructures 110
11.3 Bulk Bandstructure 111
11.4 Doped Semiconductors 112
11.5 Quantum Wells 113
11.6 Quantum Wires 114
11.7 Quantum Dots 116

12 The Ballistic Transistor 119
12.1 Introduction 119
12.2 The field-effect transistor 119
12.3 Ballistic current-voltage characteristics 123
12.4 Examples 125

13 Through the Barrier: Transmission and Tunneling 127
13.1 Tunneling and Transmission 127
13.2 The WKB Method 128
13.3 WKB Method for Semiconductor Transport 131
13.4 Tunneling transport using the WKB method 132
13.5 Interband 'Zener' Tunneling 133

14 Fermi's Golden Rule 139
14.1 Introduction 139
14.2 Fermi's Golden Rule 139
14.3 Perturbations oscillating in time 142
14.4 Transitions to a continuum of states 143
14.5 Higher order transitions: Dyson series and Diagrams 144
14.6 Fate of the initial state: Self Energy 145

15 The Boltzmann Transport Equation 147
15.1 Introduction 147
15.2 The Boltzmann Transport Equation 147
15.3 Scattering in equilibrium 149
15.4 Scattering in steady state 150
15.5 Formal Boltzmann Transport theory 151

16 Maxwell Equations in a Nutshell 155
16.1 Introduction 155
16.2 Maxwell's equations 156
16.3 Light emerges from Maxwell's equations 157
16.4 Maxwell's equations in $(\mathbf{k}, \omega)$ space 158
16.5 Maxwell's equations in material media 160
16.6 Need for a quantum theory of light 162

17 Light-Matter Interaction **165**
17.1 Introduction 165
17.2 Electron-photon matrix elements for semiconductors 165
17.3 The absorption spectrum of bulk semiconductors 167
17.4 The number of photons in light 168
17.5 Photon absorption rate in bulk semiconductors 169
17.6 The Equilibrium Absorption Coefficient $\alpha_0(\hbar\omega)$ 174

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 1

And off we go!

Contents

1.1 Beyond belief 7
1.2 A brief history of Semiconductors 7
1.3 Future 8
1.4 These boots are made for walking 9
Summary 9
Problems 9

1.1 Beyond belief

How does the smallest and lightest fundamental particle in nature - the **electron** - make it possible for me to type these words on a computer? How did the electron help light up the screen of my laptop? How did it help remember all of this text as the battery of my laptop ran out? And how did it help kick those photons that powered my email to you, and helped you download this file and read these words? Explaining this remarkable story is what we will do in this course.

The invention of the steam engine made a mockery of what was believed to be the limit of the speed at which heavy mechanical objects could be moved. In much the same way, the discovery of the electron, and in particular the discovery of the class of materials called semiconductors has made a mockery of what was believed possible for three of the deepest and most profound human endeavors: performing logic to structure and produce information (**computation**), storing information (**memory**), and transmitting and receiving information (**communication**). Our understanding of the inner workings of electrons in semiconductors has given us powers beyond belief. We depend on semiconductors today to see and talk with family and/or friends on the other side of the planet. Semiconductors empower us to generate energy from sunlight, predict weather, diagnose and treat diseases, decode the DNA, design and discover new drugs, and guide our cars in the streets as deftly as satellites in deep space. They have placed the entire recorded history of information about the universe in the palm of our hands.

In this set of notes, we will dig in to understand how all of this is possible. The semiconductor revolution is one of the most remarkable human adventures in the entire recorded history of science. The revolution has been powered by the combined efforts of scientists and engineers from several fields. The contributions of mathematicians and physicists, of chemists and materials scientists, and electrical engineers have made this possible. We all acknowledge that the artificial labeling of our skills helps with administrative purposes[1], and has no basis in science! Nowhere is this highlighted more than in the field of semiconductors in which 'engineers' have been awarded Nobel prizes in Physics, and physicists and chemists have founded successful 'engineering' semiconductor companies such as Intel on their way to becoming billionaires.

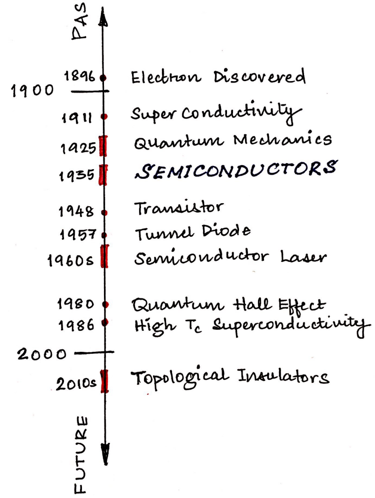


Fig. 1.1: A brief history of material properties.

1.2 A brief history of Semiconductors

Figure 1.1 shows a timeline of materials and their properties that are relevant to this course. Insulators and metals were known centuries ago. Their properties were studied, and theories were developed to explain them. In chapter 2, we will intercept this emerging

[1] Harvard's 'Division of Engineering and Applied Sciences' or DEAS changed its name to SEAS, the 'School of Engineering and Applied Sciences'. No one wants to create a division between applied sciences and engineering indeed!

story during the period right after the discovery of the electron in 1896 by J. J. Thomson. The concept of the electron subject to classical laws of mechanics, electromagnetism, and thermodynamics was used by Paul Drude to explain many properties of metals that had remained mysterious till that point of time.

Fig. 1.2: J. J. Thomson discovered the electron in 1896 @ the Cavendish Laboratory. He was awarded the 1906 Nobel prize in Physics. Seven of his students went on to win Nobel prizes.

It comes as a surprise to many that superconductivity was experimentally discovered in 1911, *decades* before semiconductors were recognized as a distinct electronic state of matter. But our understanding of the physics of semiconductors developed much more rapidly than that of superconductors, as did their applications - as highlighted in Figure 1.1. This book will make the reasons clear why this is so - that electrons in semiconductors are at first glance 'simpler' to understand because they act independently, whereas electrons in superconductors pair up and are strongly correlated. However, the simplicity of electrons in semiconductors was just a mask. Applications following the invention of the semiconductor transistor in 1948 led to very high level of control, perfection, and understanding of semiconductors, which has led to the discovery of several layers of richer physical behavior.

Esaki discovered electron tunneling in a solid - a genuine quantum mechanical effect highlighting the wave-nature of electrons - in semiconductor p-n diodes in 1957, around the same time Bardeen, Cooper, and Schrieffer proposed the eponymous BCS theory of superconductivity. The semiconductor transistor miniaturized the vaccum tube amplifier and made it far more compact, rugged, and energy efficient, setting up the stage for its use for digital logic and memory, enabling computation. In much the same way, by understanding how electrons in semiconductors interacted with photons, the invention of the semiconductor laser diode in the 1960s and 1970s shrunk the solid-state laser and revolutionized photonics, enabling semiconductor lighting and optical communications.

In 1980, von Kltizing discovered the integer quantum Hall effect while investigating electron transport in the Silicon transistor at low temperatures and at high magnetic fields. This was followed by the discovery of fractional quantum Hall effect in 1982 in high-quality quantum semiconductor heterostructures by Tsui, Stormer, and Gossard. The quantum Hall effect revealed a whole new world of condensed matter physics, because under specific constraints of lower dimensionality, and electric and magnetic fields, the behavior of electrons defied classification into any of the previously identified phases: metals, insulators, semiconductors, or superconductors. The effort to classify the quantum Hall state has led to new electronic and magnetic phases of electrons in solids: based on the *topology* of their allowed energy bands. This line of thought has led to the discovery of Topological Insulators, the newest addition that significantly enriches the field of electronic phases of matter.

The discovery of high-temperature superconductors in layered cuprates in 1987 by Bednorz and Muller again rocked the field of superconductivity, and the field of condensed matter physics. That is because the tried-and-tested BCS theory of superconductivity could not explain their behavior. At the forefront of experiments and theory today, there is a delightful connection being established between topological insulators (that have emerged from semiconductors), and superconductivity (that have emerged from metals). There are now proposals for topological superconductivity, and semiconductor - superconductor heterostructures that can make electrons behave in much stranger ways - in which they pair up to lose their charge and spin identities and can 'shield' themselves from electric and magnetic fields. Such particle-antiparticle pairs, called Majorana Fermions can enable robust quantum-bits (qubits) for quantum computation in the future. If it sounds like science fiction[2] - let me assure you it is not, the race is on to experimentally find these elusive avatars of the electron in many laboratories around the world!

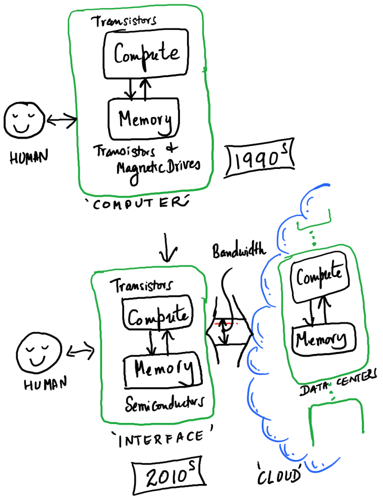


Fig. 1.3: How semiconductor technologies have changed the handling of data and information over the last two decades.

1.3 Future

Semiconductors today power the 'information age' in which data is produced in high volumes, and transported at the speed of light. Figure 1.3 shows the underlying structure.

[2] The 2009 movie Avatar showed humans trying to mine the fictional material 'Unobtanium' from Pandora - a room-temperature superconductor.

In the 1990s, most computers were stand-alone devices that we interacted with and programmed directly. They boasted semiconductor integrated circuit-based microprocessors for computation, and semiconductor based random-access memories (RAMs), and magnetic (or spin-based) hard drives or Read-Only Memories (ROMs).

Today, a large and increasing part of the computing infrastructure resides in the 'cloud'. The cloud is loosely also made of a large and distributed network of microprocessors and memories that may be housed in server farms. Much of the heavy computational tasks and memory storage is thus not performed on the computers in front of us with which we directly interact. The communication links that connect our interface devices with the cloud are therefore increasingly important, and their bandwidth will determine the efficiency of such networks. The computational and memory capacity created by miniaturization of transistors and memory elements, and large bandwidth communication links powered by semiconductor lasers has also led, in part to an explosion of social media.

The first semiconductor 'point-contact' transistor of 1947 was a few centimeters in size. Figure 1.4 shows this device in comparison to a much newer transistor - the Silicon finFET of 2012. The transistor is so small today that you can count the number of atoms across the fin - the scale has shrunk by a factor 10^7 from centimeter in 1947 to nanometer in 2012. In the latest generation microprocessors, several billion transistors can fit into a cm-size chip. The orders-of-magnitude increase in computational power and storage capability is a direct result of this miniaturization. Such orders of magnitude improvements over sustained periods of time is extremely rare in the recorded history of science.

Where are things going next? Some semiconductor experts and accomplished practitioners in this art will tell you that the golden era of semiconductors is over. That transistors have reached such small dimensions and lasers are so efficient that there is not much more one can do. Do not buy into such predictions, because most such predictions can be proven to be wrong invoking physics[3]. Go back and read this chapter once more. Just like it has happened again and again, some of you will do experiments or come up with a theory that will make a mockery of all current belief of what is possible with semiconductors and related materials. The future of our civilization depends on this adventure that you must embark upon! This book is an effort to motivate you to take this journey by arming you with the science that will serve as your fuel.

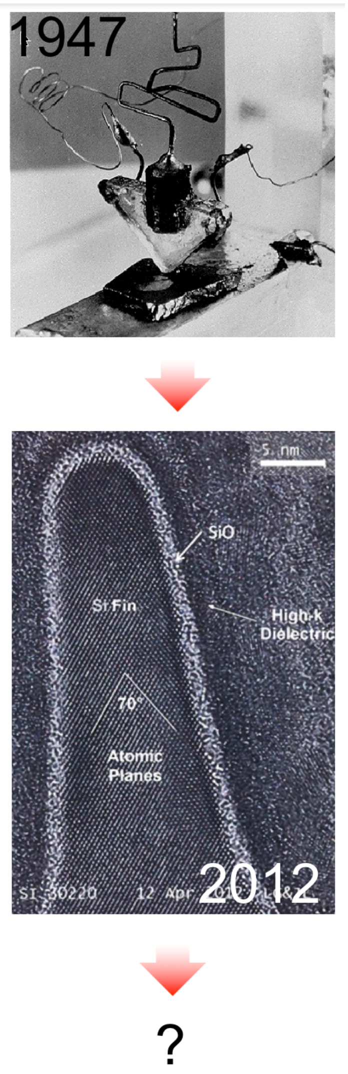


Fig. 1.4: The first centimeter-scale point-contact transistor, and today's nanometer-scale FinFETs.

1.4 These boots are made for walking

This book is meant to be read with enjoyment and wonder, the concepts discussed in and outside of class, in office hours, and over coffee. Every effort will be made to present the development of the field as a uniquely human endeavor - laws and theorems did not fall out of the sky, but are creations of precise experimental observation and recognition of mathematical patterns - both achieved by human ingenuity. Each chapter is of a length that should take one sitting to complete reading. Your feedback is most useful in making sure that these lofty goals are indeed achieved.

Chapter Summary

At the end of each chapter we will summarize the major ideas in a few sentences.

Problems

[3] "It is exceedingly difficult to make predictions, especially about the future" - Danish proverb, sometimes attributed to Niels Bohr invoking the uncertainty principle in his answer to the question - "What will be the impact of Quantum Mechanics on the future of the world?"

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 2

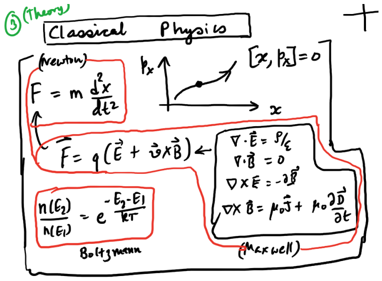
Secrets of the classical electron

Contents

2.1 Our ancestors knew metals 11
2.2 Discovery of the electron and its aftermath 11
2.3 Drude's model explains Ohm's law 12
2.4 Metals are Shiny 13
2.5 Metals conduct heat . . 13
2.6 Icing on the cake: The Weidemann-Franz Law 14
2.7 All is *not* well 14
Summary 15
Problems 16

I would like to emphasize from the beginning how experiments have driven the search for new theories. New theories are born when experimental facts defy the sum total of all existing theories. This happens again and again because when you perform an experiment, you cannot stop a physical law from doing its thing. Meaning, any experiment ever performed has measured every *known* law of physics, and every *yet to be discovered* law of physics! So when you measure something in regimes no one has ever done before, because of better equipment, or because you are just plain clever, new physical phenomena reveal themselves to the careful researcher.

We will repeatedly observe this dynamics in play. In this handout we glimpse how the pre-quantum era explained the observed properties of metals using the three pillars on which physics rested around 1900s: Newton's **Classical Mechanics**, Maxwell's **Electromagnetism**, and Boltzmann's **Thermodynamics**. But then clever experimentalists pushed measurements to new regimes which revealed physical phenomena that were in stark contradiction to these three theories. Let's start talking about metals.



3 (Theory)
Classical Physics
(Newton)
$F = m\frac{d^2x}{dt^2}$
p_x
x
$[x, p_x] = 0$
$\vec{F} = q(\vec{E} + \vec{v}\times\vec{B})$
$\nabla\cdot\vec{E} = \rho/\epsilon$
$\nabla\cdot\vec{B} = 0$
$\nabla\times\vec{E} = -\partial\vec{B}$
$\nabla\times\vec{B} = \mu_0\vec{J} + \mu_0\frac{\partial\vec{D}}{\partial t}$
$\frac{n(E_2)}{n(E_1)} = e^{-\frac{E_2-E_1}{kT}}$
Boltzmann
(Maxwell)

Fig. 2.1: The three pillars of classical physics

2.1 Our ancestors knew metals

Because a large number of elements in the periodic table are metals and many occur in nature, they were discovered early, way before 1900s. Metals have been known for a very long time to be very different from insulators by being[1]

- good conductors of electricity,
- good conductors of heat, and
- reflective and shiny.

Wiedemann and Franz in 1850s had even discovered a deep connection between the thermal conductivity κ and electrical conductivity σ of metals: they had found the quantity $\frac{\kappa}{\sigma T} \sim 10^{-8}(\frac{\mathrm{V}}{\mathrm{K}})^2$, where T is the temperature, to be constant[2] for many different metals. But why? Though there were several attempts to explain all these physical characteristics of metals, a truly satisfying explanation had to wait for the discovery of the single entity responsible for each of these three features: the **electron**.

2.2 Discovery of the electron and its aftermath

J. J. Thomson discovered the electron[3] in 1897 in the Cavendish laboratory. He found the existence of a particle of mass $m_e = 9.1 \times 10^{-31}$ kg and electrical charge $q = 1.6 \times 10^{-19}$ Coulomb. This was followed by the discovery of the nucleus by Ernest Rutherford in 1911,

[1] All these properties were known before the discovery of the electron, or the atom.

[2] This experimental fact is the empirical Wiedemann-Franz law.

[3] The word 'electron' represents an indivisible portion of 'electricity' - the word electricity predates the word electron by several centuries. It motivated a number of '..on' names: photon, proton, neutron, phonon, fermion, boson...

and the neutron by James Chadwick way later in 1932. In this chapter, the nucleus of the atom will play a cameo role of passive obstructions in the path of freely moving electrons in a metal[4]. It will increasingly assert its role in later chapters.

2.3 Drude's model explains Ohm's law

The discovery of the electron was the trigger that precipitated an explanation of most properties of the metal discussed in section 2.1. This was achieved by Paul Drude , who combined the notion that electrons move in a metal just like molecules move in a gas[5] following the laws of classical mechanics, thermodynamics, and electromagnetism.

The first task was to explain the electrical conductivity of metals. The experimental fact measured for metals is that the charge current I is linearly proportional to the voltage V, stated as the Ohm's law $V = IR$, where R is the *resistance* of the metal. Drude imagined that the metal is filled with electrons of volume density in cm^{-3} units $n = \frac{N}{V}$, where N is the total number of electrons in the metal, and $V = AL$ is the volume as indicated in Figure 2.4. The current density J measured in Amp/cm^2 is defined by $I = JA$. The expression for the current density is $\boxed{J = qnv}$.

$q \rightarrow$: The charge current density is given by the flux density times the charge. The flux density of particles is the volume density times the velocity, $n \times v$. Each electron in this flux drags along with it a charge of q, so the current density is $J = qnv$. Later when we encounter heat, spin, or other currents, we will simply multiply the particle flux density with the corresponding quantity that is dragged along by the particle.

$n \rightarrow$: Because the structure of the atom was not known at the time, Drude's electron density n is an empirical number, of the order $10^{23}/\text{cm}^3$. We will see later that metals have *much more* electrons - most electrons however are stuck in core states or filled bands - which require quantum mechanics to explain. But some are free to wander around the crystal and conduct charge current - those are the *conduction electrons* Drude's quantity n considers, the others simply do not play a role.

$v \rightarrow$: Drude found the velocity of the electrons in the metal with the following argument. The force on the electrons is due to the **electromagnetic**, or Lorentz force $\mathbf{F} = q\mathbf{E} + \mathbf{v} \times \mathbf{B}$. If $\mathbf{B} = 0$, let's assume the scalar form for simplicity where $E = V/L$ is the electric field exerted on the electrons by the battery. Because of this force, electrons accelerate according to Newton's law $F = \frac{dp}{dt}$, where $p = m_e v$ is the electron momentum. But as they speed up, sooner or later they will bump into the atoms in the metal, just as a swarm of bees drifting blindly through a forest. Whenever such a collision occurs, the electron will lose all its momentum, and start out from zero momentum. The modified Newton's law taking into account the *dissipation* or damping of momentum every τ seconds is:

$$qE = m_e \frac{dv}{dt} - \frac{m_e v}{\tau}, \tag{2.1}$$

where the increase in momentum due to the force is tempered by the decrease upon collisions every τ seconds. If we ask what happens in the 'steady state', meaning we have applied the voltage and waited long enough that all transients have died out, we can use $\frac{d}{dt}(...) \rightarrow 0$, which yields the velocity v as

$$v = -\frac{q\tau}{m_e} E = \mu E \implies \boxed{\mu = \frac{q\tau}{m_e}}. \tag{2.2}$$

The electrons achieving a steady state velocity in the presence of a constant force is similar to a parachutist reaching a terminal velocity in spite of the constant gravitational force. The drift velocity is proportional to the electric field, and the constant of proportionality $\mu = \frac{q\tau}{m_e}$ is defined as the **mobility** of electrons. The concept of mobility will prove to be more useful for semiconductors than metals, as we will see later. It is clear that if the electron scatters less often, $\tau \uparrow \implies \mu \uparrow$.

Fig. 2.2: Paul Drude in 1900 proposed a model that combined classical mechanics, electromagnetism, and thermodynamics to explain the may properties of metals by invoking the then newly-discovered electron.

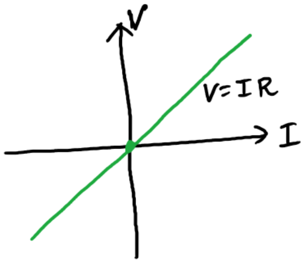


Fig. 2.3: Ohm's law is $V = IR$, or equivalently $J = \sigma E$.

[4] Protons help maintain charge neutrality, and the neutrons stabilize the nucleus and keeps the atom from disintegrating.

[5] Hence the name 'electron gas'.

Now putting together the all the above pieces, Drude found that the current density is

$$J = qnv = \frac{nq^2\tau}{m_e}E = \sigma E \implies \boxed{\sigma = \frac{nq^2\tau}{m_e}}, \tag{2.3}$$

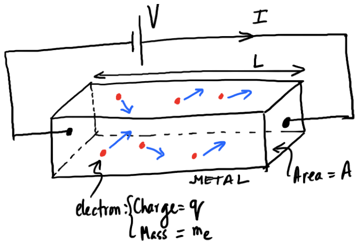


Fig. 2.4: Electron gas moving in response to an electric field in a metal.

where it is seen that the current density is proportional to the electric field, with the proportionality constant $\sigma = \frac{nq^2\tau}{m_e}$ called the **conductivity**. The components appeal to our intuition - more electrons, and longer scattering times should lead to higher conductivity. The conductivity does not depend on the sign of the charge.

To close the story, we revert back to the current and see how Drude's model explained Ohm's law of the electrical conductivity of metals:

$$I = JA = \sigma EA = \sigma\frac{V}{L}A \implies V = I\cdot\frac{1}{\sigma}\frac{L}{A} = IR \implies \boxed{R = \frac{1}{\sigma}\frac{L}{A} = \rho\frac{L}{A}}. \tag{2.4}$$

The resistance R is measured in Ohms. It is related to the microscopic details provided by Drude via the conductivity σ or the resistivity $\rho = \frac{1}{\sigma}$, and to the macroscopic dimensions[6] via L and A. A longer metal has more resistance because it takes an electron longer to get through. This is the essence of classical mechanics applied to the electron.

2.4 Metals are Shiny

Why do metals reflect light? The secret lies in the swarm of conduction electrons in them! Maxwell had shown that a beam of light is a wave of oscillating electric and magnetic fields. The electric field of a light beam is $E(t) = E_0 e^{i\omega t}$, where the circular frequency $\omega = ck = c\frac{2\pi}{\lambda} = 2\pi f$ is linked to the speed of light c and its wavelength λ. If we subject the swarm of electrons in the metal not to the constant DC voltage of a battery as we did in section 2.3, but to this new pulsating field, we can explain why metals reflect light. This will be done in Problem **??**.

Fig. 2.5: James Clerk Maxwell in 1865 unified electricity and magnetism. By introducing the concept of the displacement current, he showed that light is an electromagnetic wave. He made significant contributions to several other fields of physics and mathematics.

2.5 Metals conduct heat

Drude's model of electrical conductivity leads naturally to an explanation of the thermal conductivity of metals. Atoms likely play a small role in the thermal conductivity because the density of atoms are similar in metals and insulators. Because typical electrical insulators are also poor thermal conductors[7], the thermal conductivity of metals must be due to the conduction electrons.

Based on this reasoning, a model for the thermal conductivity of metals due to electrons alone goes like this: consider a piece of metal as shown in Figure 2.6. The volume density of electrons $n = \frac{N}{V}$ is uniform across the metal, but the left end is held at a hotter temperature T_1 than the right end, which is at T_2. There is a temperature gradient across the metal from the left to the right. Drawing analogy to the charge current density $J_{charge} = \sigma\cdot(-\nabla V)$, where the potential gradient $-\nabla V = E$ is the electric field, and σ the electrical conductivity, we hunt for an expression for the heat current in the form $J_{heat} = \kappa\cdot(-\nabla T)$, where ∇T is the temperature gradient. If we can write the heat current in this form, we can directly read off the thermal conductivity κ. Because the heat current density J_{heat} has units $\frac{\text{J}}{\text{cm}^2\text{s}} = \frac{\text{Watt}}{\text{cm}^2}$, the thermal conductivity κ must have units of $\frac{\text{Watts}}{\text{cm}\cdot\text{K}}$.

Looking at Figure 2.6, we zoom into the plane x, and ask how much heat energy current is whizzing past that plane to the right. Because the density of conduction electrons is uniform, the heat current is flowing because of a difference in energy via the temperature gradient. Let $\mathcal{E}$ be the individual electron energy. The energy depends on x via the temperature at that plane. In the plane one mean free path $v_x\tau$ to the left of x, the energy

[6] This concept works well for large resistors. But we will see later that this picture will fail spectacularly when the L and A become very small, comparable to the wavelength of electrons.

[7] Diamond, BN, SiC, and AlN are exceptions, to be discussed later.

of electrons is $\mathcal{E}[T(x - v_x\tau)]$. Electrons in this (hotter) side have energies higher than those one mean-free path to the right of plane x, which have the energy $\mathcal{E}[T(x + v_x\tau)]$. Half the carriers at $x - v_x\tau$ are moving to the right, carrying a heat current density $\frac{n}{2}v_x\mathcal{E}[T(x - v_x\tau)]$. Similarly, half the carriers at $x + v_x\tau$ transport heat current $\frac{n}{2}v_x\mathcal{E}[T(x + v_x\tau)]$ to the left. The net heat current at x is the current flowing to the right, minus the current flowing to the left:

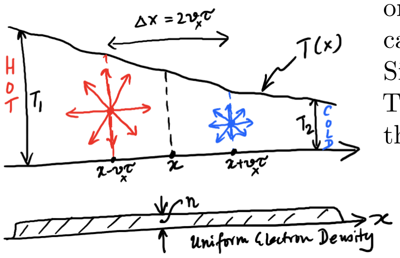


Fig. 2.6: Figure showing how a uniform density of electrons n in a metal can transport heat energy from the hot to the cold side.

$$J_{heat} = \frac{n}{2}v_x[\mathcal{E}[T(x - v_x\tau)] - \mathcal{E}[T(x + v_x\tau)]]. \tag{2.5}$$

Now if we assume that the mean free paths $v_x\tau$ are small, we can write the heat current as

$$J_{heat} = \frac{n}{2}v_x\frac{\Delta\mathcal{E}}{\Delta T}\frac{\Delta T}{\Delta x}\Delta x = \frac{n}{2}v_x\frac{d\mathcal{E}}{dT}\frac{dT}{dx}(-2v_x\tau) \implies \boxed{J_{heat} = \frac{1}{3}c_v v^2\tau(-\nabla T)}. \tag{2.6}$$

In writing the final boxed version of equation 2.6, we have used the fact that because of motion of electrons in 3 dimensions, $v_x^2 = \frac{v^2}{3}$, and the electron heat capacity is given by $c_v = \frac{1}{V}\frac{d\mathcal{U}}{dT}$, where $\mathcal{U} = N\mathcal{E}$ is the total energy of the electron system, and $n = \frac{N}{V}$.

Thus, the Drude model of heat current carried by electrons in a metal gives us a thermal conductivity $\kappa = \frac{1}{3}c_v v^2\tau$ by analogy to the charge current. But the Drude model did more: it also was able to give an explanation of the ratio of electrical and thermal conductivity.

2.6 Icing on the cake: The Weidemann-Franz Law

One of the major successes of the Drude electron model of electrical and thermal conductivity of metals achieved was to provide an explanation for the Wiedemann-Franz law, which had languished for half a century without an explanation.

From Equations 2.3 and 2.6, we get the ratio

$$\frac{\kappa}{\sigma T} = \frac{(\frac{1}{3}c_v v^2\tau)}{(\frac{nq^2\tau}{m_e})T} = \frac{(\frac{1}{3}\frac{3}{2}nk_B\frac{3k_BT}{m_e}\tau)}{(\frac{nq^2\tau}{m_e})T} = \frac{3}{2}(\frac{k_B}{q})^2 \implies \boxed{\frac{\kappa}{\sigma T} = \frac{3}{2}(\frac{k_B}{q})^2 = \mathcal{L}}. \tag{2.7}$$

Here we have invoked the classical kinetics result that the heat capacity of electrons is $c_v = \frac{1}{V}\frac{d\mathcal{U}}{dT} = \frac{1}{V}\frac{d(N\cdot\frac{3}{2}k_BT)}{dT} = \frac{3}{2}nk_B$, and the velocity is obtained from the thermal energy by the equipartition relation: $\frac{1}{2}m_e v^2 = \frac{3}{2}k_BT$. Every microscopic detail specific to a particular metal cancels out in the ratio above, and what remains are the two constants that underpin classical physics: the Boltzmann constant k_B, and the electron charge q. This seemed to explain the Weidemann-Franz law beautifully. The ratio is called the Lorenz number $\mathcal{L}$ with a value $\sim 10^{-8}(\frac{\mathrm{V}}{\mathrm{K}})^2$. Thus, Drude's model of a classical electron gas seems to have resolved the mystery we set out with in section 2.1.

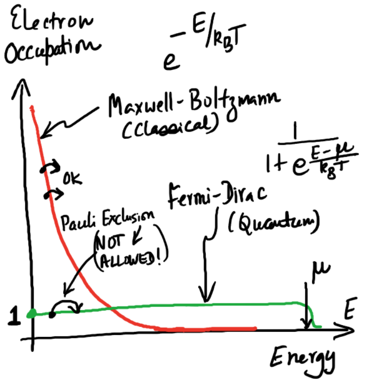
2.7 All is *not* well

The success of the Drude model to explain the Weidemann-Franz law is remarkable, but unfortunately it is fundamentally flawed. We will see later that there is a crucial cancellation of two unphysical quantities that leaves the ratio intact. When experimentalists measured the specific heat of electrons in metals, it was found to be *much smaller* than the classical result $c_v = \frac{3}{2}nk_B$ which was used by Drude. The value was found to be *much smaller*. This is a null result[8] that hints at something deeper - and we will see soon that something is quantum mechanics.

[8] Huxley: The great tragedy of science: the slaying of a beautiful hypothesis by an ugly fact.

We will see in subsequent chapters that by demanding that electrons follow the rules of quantum mechanics instead of Newtonian mechanics, the electronic heat capacity was shown by Arnold Sommerfeld to be $c_v = [\frac{\pi^2}{2} n k_B] \cdot [\frac{k_B T}{E_F}]$ instead of the classical $c_v = \frac{3}{2} n k_B$. Here $E_F = \frac{p_F^2}{2m_e}$ is the **Fermi Energy**, $p_F = \hbar k_F$ the **Fermi Momentum**, and $k_F = (3\pi^2 n)^{\frac{1}{3}}$ is the **Fermi Wavevector**. These three quantities are quantum-mechanical concepts, and simply cannot be explained by classical physics. Aside from the constants in the correct expression for the electron heat capacity, only a *small fraction* of the n conduction electrons: $\frac{k_B T}{E_F}$ to be precise, seem to actually contribute to the heat capacity. We now know the reason: electrons being fermions, are subject to the **Pauli Exclusion Principle**. This principle prevents two electrons from occupying the same state in the metal.

The consequence of the exclusion principle is rather drastic on the distribution of electrons. Figure 2.7 highlights this difference. In classical statistical physics, the number of electrons at an energy E goes as $e^{-E/k_B T}$, which is the Maxwell-Boltzmann distribution. So most electrons will pile up at the lowest allowed energies $E \to 0$ to lower the system energy. When extra energy is pumped into the electron system by whatever means, electrons at all energies in the distribution have equal opportunity to increase their energy. This is what led to the classical heat capacity $c_v = \frac{1}{V} \frac{d(N \times \frac{3}{2} k_B T)}{dT} = \frac{3}{2} n k_B$.



Electron Occupation
e^{-E/k_BT}
Maxwell-Boltzmann (Classical)
OK
Pauli Exclusion (NOT ALLOWED!)
$\frac{1}{1+e^{\frac{E-\mu}{k_BT}}}$
Fermi-Dirac (Quantum)
μ
1
E
Energy

Fig. 2.7: Figure showing how the much higher heat capacity of electrons comes about because of assuming the classical Maxwell-Boltzmann distribution in energy. The Pauli exclusion principle of quantum mechanics changes the distribution to the Fermi-Dirac distribution, which fixes the electronic specific heat anomaly of the Drude model.

However, when electrons are forced to obey the Pauli exclusion principle, once an allowed state at a low energy is occupied by an electron, it makes it impossible for a second electron to occupy the same state. The second electron must occupy a higher energy state. If we continue filling the states till the last electron, the highest energy occupied is referred to as the **chemical potential** μ. At this stage, we will assume that the chemical potential $\mu = E_F$ is equal to the Fermi energy; in later chapters we will discuss their differences. The occupation probability of a state of energy E upon enforcing the Pauli Exclusion Principle is the Fermi-Dirac distribution, $f(E) = \frac{1}{1+e^{\frac{e-\mu}{k_B T}}}$, whose maximum value is 1. This is shown in Figure 2.7. We will discuss this distribution in significant detail later.

The Fermi-Dirac distribution of electrons makes it clear why only a fraction $\frac{k_B T}{E_F}$ of electrons in a metal can actually absorb energy and promote themselves to higher energy states. Because of the Pauli exclusion principle, none of the electrons at energies from $0 < E < E_F - 3k_B T$ can increase their energy by absorbing $k_B T$ energy, because they are Pauli-blocked. Electrons in only a tiny sliver of energies $k_B T$ around the Fermi energy E_F have the freedom to jump to higher energy states that are unoccupied. Thus, the electronic heat capacity c_v is much smaller than what was predicted by the Drude model. In the next chapter, we discuss the basics of quantum mechanics and see why electrons must follow the Fermi-Dirac distribution in the first place.

Chapter Summary

- The electronic and thermal conductivities of metals were explained by the Drude model by attributing these properties correctly to the newly discovered particle, the electron.

- In the Drude model, free electrons subject to the laws of classical mechanics, electromagnetism, and thermodynamics could explain the electronic conductivity σ, and the thermal conductivity κ reasonably successfully.

- The Drude model also seemed to resolve a long-standing mystery of the empirical Wiedemann-Franz law, which stated the ratio $\frac{\kappa}{\sigma T}$ is a constant for metals.

- The heat capacity of free conduction electrons in metals predicted by the Drude model turned out to be inconsistent with the measured values, which were several orders too small. It would need the full machinery of quantum mechanics and a quarter century to resolve this discrepancy.

Problems

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 3

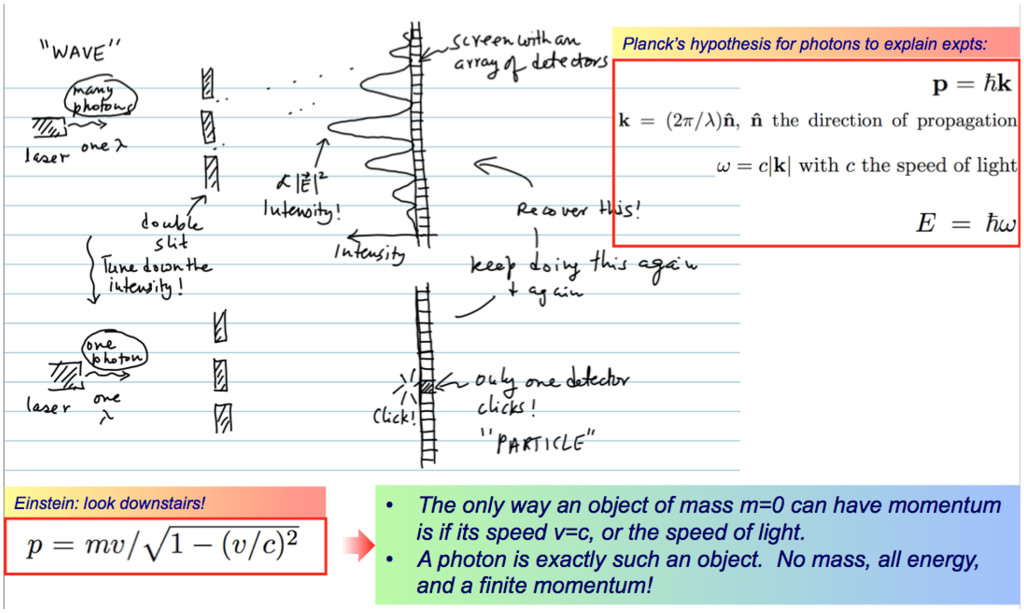
Quantum Mechanics in a Nutshell

Contents

3.1 Photons 17
3.2 Wave-Particle Duality . . 18
3.3 The wavefunction 19
3.4 Operators 20
3.5 States of definite momentum and location . 21
3.6 States of definite energy: The Schrodinger equation 23
3.7 Time-dependent Schrodinger equation 23
3.8 Stationary states and time evolution 25
3.9 Quantum Current 25
3.10 Fermions and Bosons . . . 27
3.11 Spin, and the Spin-Statistics Theorem 29
3.12 The Dirac equation and the birth of particles . 29
Summary 30
Problems 30

This chapter presents a very short summary of the major ideas of quantum mechanics. By tracing the historical development of the ideas, we learn how we have learnt to treat the dynamics of electrons by the laws of quantum mechanics rather than Newton's laws. It is highly recommended that you append your reading of this chapter with your favorite quantum mechanics texts.

3.1 Photons



Planck's hypothesis for photons to explain expts:

$$\mathbf{p} = \hbar\mathbf{k}$$

$\mathbf{k} = (2\pi/\lambda)\hat{\mathbf{n}}$, $\hat{\mathbf{n}}$ the direction of propagation

$\omega = c|\mathbf{k}|$ with c the speed of light

$$E = \hbar\omega$$

Einstein: look downstairs!

$$p = mv/\sqrt{1-(v/c)^2}$$

- The only way an object of mass m=0 can have momentum is if its speed v=c, or the speed of light.
- A photon is exactly such an object. No mass, all energy, and a finite momentum!

Fig. 3.1: Photons behaving as particles.

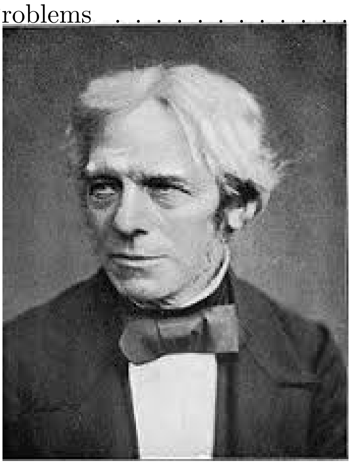


Fig. 3.2: Michael Faraday, considered to be one of the greatest experimentalists of all times. Discovered the relation between electric and magnetic fields, and influenced Maxwell to discover that light is an electromagnetic wave. Light played *the* central role in the development of quantum mechanics.

Time: end of the 19th century. Maxwell's equations have established Faraday's hunch that light is an electromagnetic wave. However, by early 20th century, experimental evidence mounted pointing towards the fact that light is carried by 'particles' that pack a definite momentum and energy. Here is the crux of the problem: consider the double-slit experiment. Monochromatic light of wavelength λ passing through two slits separated by a distance $d \sim \lambda$ forms a diffraction pattern on a photographic plate. If one tunes down

Fig. 3.3: Max Planck, the ‘father’ of quantum mechanics. Postulated quanta of light (photons) to explain the blackbody radiation spectrum. Was awarded the Nobel prize in 1918.

the intensity of light in a double-slit experiment, one does not get a ‘dimmer’ interference pattern, but discrete strikes on the photographic plate and illumination at specific points. That means light is composed of ‘particles’ whose energy and momentum are concentrated in one point which leads to discrete hits. But their wavelength extends over space, which leads to diffraction patterns.

Planck postulated that light is composed of discrete lumps of momemtum $\mathbf{p} = \hbar\mathbf{k}$ and energy $E = \hbar\omega$. Here $\mathbf{k} = (2\pi/\lambda)\hat{\mathbf{n}}$, $\hat{\mathbf{n}}$ the direction of propagation, $\hbar$ is Planck's constant, and $\omega = c|\mathbf{k}|$ with c the speed of light. Planck's hypothesis explained spectral features of the blackbody radiation. It was used by Einstein to explain the photoelectric effect. Einstein was developing the theory of relativity around the same time. In this theory, the momentum of a particle of mass m and velocity v is $p = mv/\sqrt{1-(v/c)^2}$, where c is the speed of light. Thus if a particle has $m = 0$, the only way it can pack a momentum is if its velocity is $v = c$. Nature takes advantage of this possibility and gives us such particles. They are now called photons. Thus photons have no mass, but have momentum. Thus light, which was thought a wave acquired a certain degree of particle attributes. So what about particles with mass - do they have wave nature too? Nature is too beautiful to ignore this symmetry!

3.2 Wave-Particle Duality

de Broglie hypothesized in his PhD dissertation that classical ‘particles’ with mass also have wavelengths associated with their motion. The wavelength is $\lambda = 2\pi\hbar/|\mathbf{p}|$, which is identical to $\mathbf{p} = \hbar\mathbf{k}$. How could it be proven? The wavelength of light was such that diffraction gratings (or slits) were unavailable at that time. But electron wavelengths were much shorter, since they had substantial momentum due to their mass. Elsassaer proposed using a crystal where the periodic arrangement of atoms will offer a diffraction grating for electrons. Davisson and Germer at Bell labs shot electrons in a vacuum chamber on the surface of crystalline Nickel. They observed diffraction patterns of electrons. The experiment proved de Broglie's hypothesis was true. All particles had now acquired a wavelength.

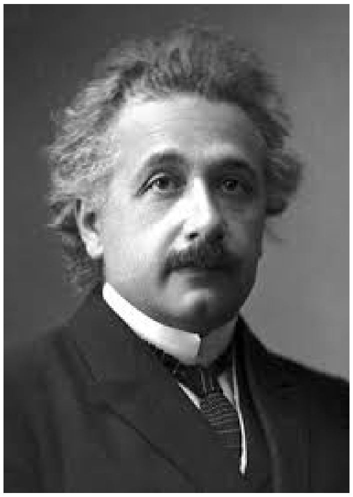


Fig. 3.4: Albert Einstein is considered the greatest physicist since Newton. In addition to special and general relativity, contributed significantly to the development of quantum mechanics, and all areas of physics. Nobel prize in 1921 for the photoelectric effect, an early experiment confirming quantum theory. Did not have a brother, and certainly did *not* start a bagel company.

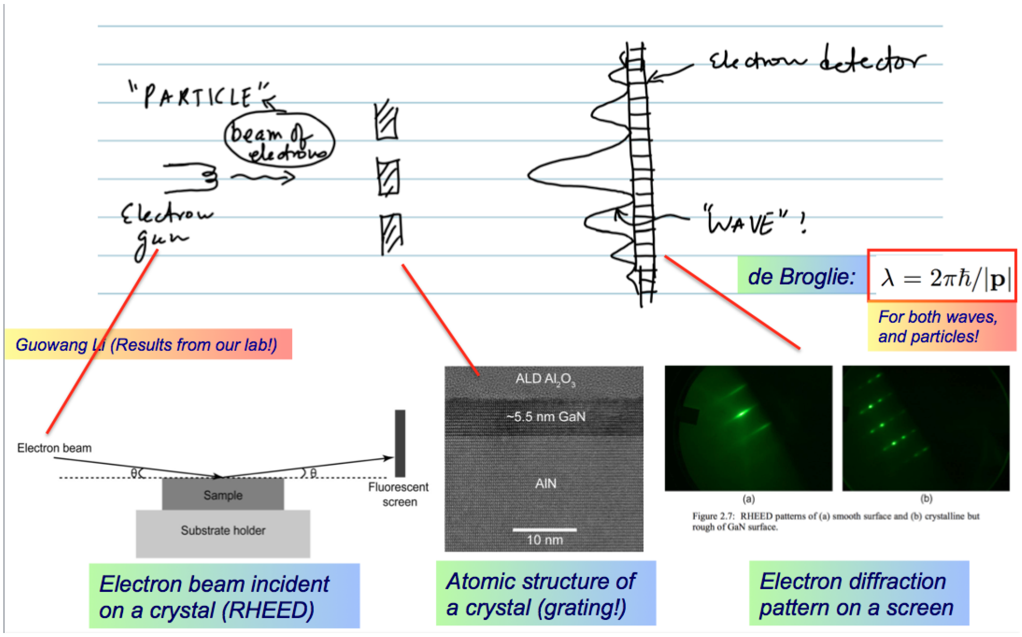

"PARTICLE"
beam of electrons
Electron gun
Electron detector
"WAVE"!
de Broglie: $\lambda = 2\pi\hbar/|\mathbf{p}|$
For both waves, and particles!
Guowang Li (Results from our lab!)
Electron beam
θ
θ
Sample
Substrate holder
Fluorescent screen
ALD Al_2O_3
~5.5 nm GaN
AlN
10 nm
(a)
(b)
Figure 2.7: RHEED patterns of (a) smooth surface and (b) crystalline but rough of GaN surface.
Electron beam incident on a crystal (RHEED)
Atomic structure of a crystal (grating!)
Electron diffraction pattern on a screen

Fig. 3.7: Electrons behaving as waves.

The experiment challenged the understanding of the motion or 'mechanics' of particles, which was based on Newton's classical mechanics. In classical mechanics, the question is the following: a particle of mass m has location x and momentum p now. If a force F acts on it, what are (x', p') later? Newton's law $F = md^2x/dt^2$ gives the answer. The answer is deterministic, the particle's future fate is completely determined from its present. This is no longer correct if the particle has wave-like nature. The wave-particle duality is *the central fabric* of quantum mechanics. It leads to the idea of a wavefunction.

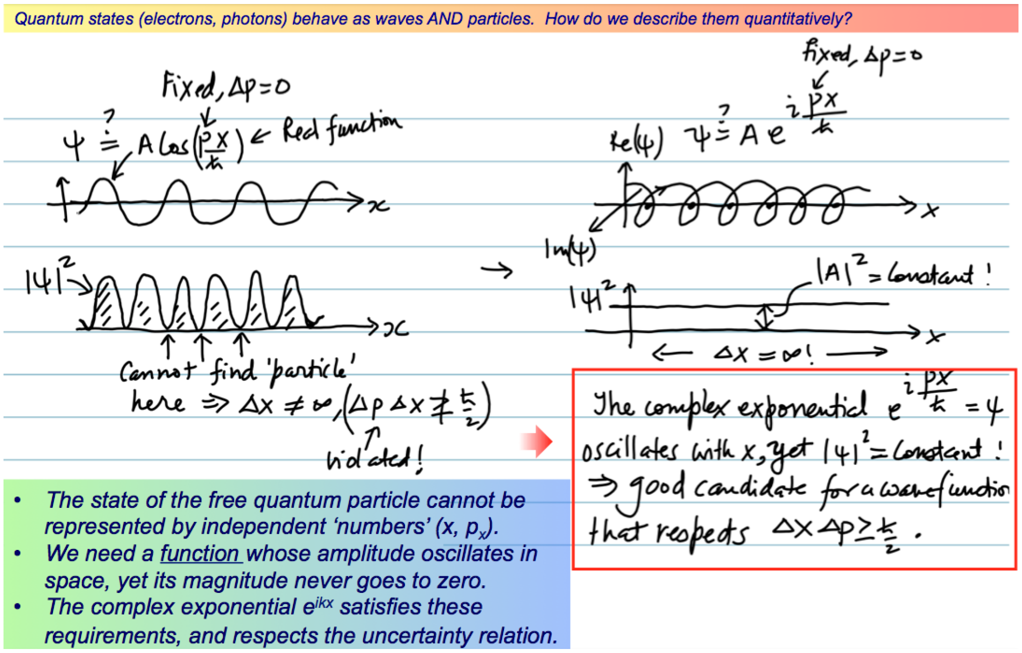
Fig. 3.5: de Broglie proposed in his PhD thesis that particles with mass have wavelengths associated with their motion. Was awarded the Nobel prize in Physics in 1929.

3.3 The wavefunction

If a particle has a wavelength, what is its location x? A wave is an extended quantity. If a measurement of the particle's location is performed, it may materialize at location x_0. But repeated measurements of the same state will yield $\langle x \rangle = x_0 + \Delta x$. Separate measurements of the momentum of the particle prepared in the same state will yield $\langle p \rangle = p_0 + \Delta p$. The 'uncertainty' relation $\Delta x \Delta p \geq \hbar/2$ is a strictly mathematical consequence of representing a particle by a wave.

Because the 'numbers' (x, p) of a particle cannot be determined with infinite accuracy simultaneously, one has to let go of this picture. How must one then capture the mechanics of a particle? Any mathematical structure used to represent the particle's state must contain information about its location x *and* its momentum p, since they are forever intertwined by the wave-particle duality. One is then forced to use a *function*, not a number. The function is denoted by ψ, and is called the wavefunction.

Fig. 3.6: Joseph Fourier is the best known French mathematician and physicist. Good friend of Napoleon.



Quantum states (electrons, photons) behave as waves AND particles. How do we describe them quantitatively?

Fixed, $\Delta p = 0$
$\psi \overset{?}{=} A\cos\left(\frac{px}{\hbar}\right)$ ← Real function
x
$|\psi|^2$
x
Cannot find 'particle' here ⇒ $\Delta x \neq \infty$, ($\Delta p \Delta x \ngeq \frac{\hbar}{2}$)
violated!
→
fixed, $\Delta p = 0$
$\mathrm{Re}(\psi)$ $\psi \overset{?}{=} A e^{i\frac{px}{\hbar}}$
$\mathrm{Im}(\psi)$
x
$|\psi|^2$
$|A|^2 = $ Constant !
x
← $\Delta x = \infty$! →

The complex exponential $e^{i\frac{px}{\hbar}} = \psi$ oscillates with x, yet $|\psi|^2 = $ Constant ! ⇒ good candidate for a wavefunction that respects $\Delta x \Delta p \geq \frac{\hbar}{2}$.

- *The state of the free quantum particle cannot be represented by independent 'numbers' (x, p_x).*
- *We need a function whose amplitude oscillates in space, yet its magnitude never goes to zero.*
- *The complex exponential e^{ikx} satisfies these requirements, and respects the uncertainty relation.*

Fig. 3.10: Birth of the wavefunction to account for the wave-particle duality.

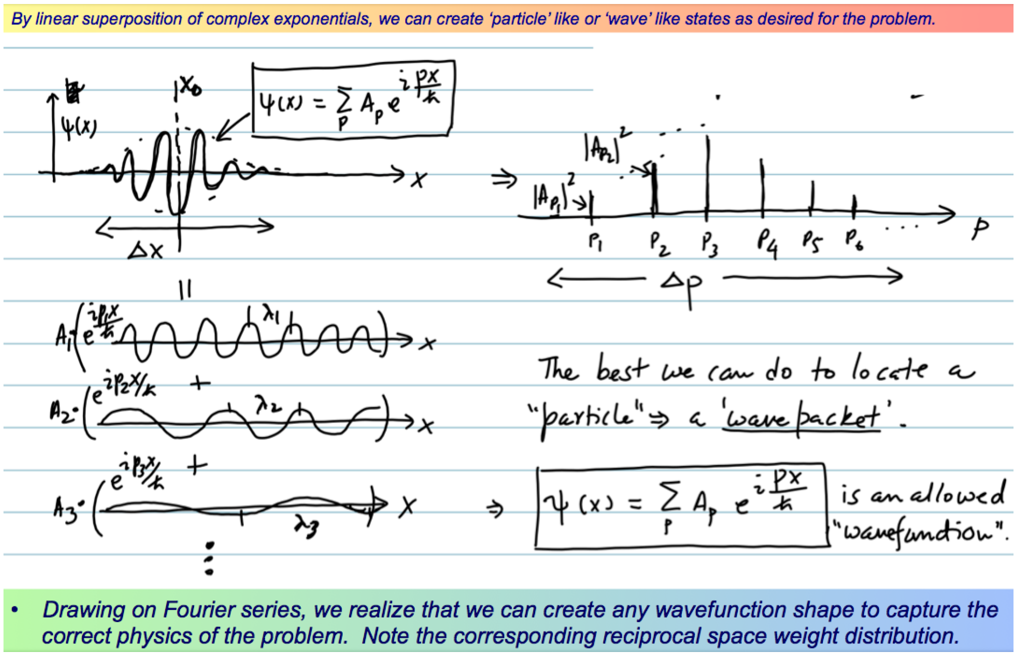
A first attempt at constructing such a function is $\psi(x) = A\cos(px/\hbar)$. This guess is borrowed from the classical representation of waves in electromagnetism, and in fluid dynamics. The wavefunction can represent a particle of a definite momentum p. Max Born provided the statistical interpretation of the wavefunction by demanding that $|\psi|^2$ be the probability density, and $\int |\psi|^2 dx = 1$. In this interpretation, $|\psi(x)|^2 dx$ is the probability that a measurement of the particle's location will find the particle in the location $(x, x+dx)$. It is clear that $|\psi(x)|^2 = |A|^2 \cos^2(px/\hbar)$ assigns specific probabilities of the location of the

Fig. 3.8: Max Born introduced the probabilistic representation of the quantum wavefunction. With his student Heisenberg, discovered matrix mechanics. Influenced and guided several young scientists who made contributions to quantum theory. Nobel prize 1954.

Fig. 3.9: Werner Heisenberg discovered matrix mechanics with Max Born, and is an original founder of quantum theory. Nobel prize in 1932. Better known in pop media for his uncertainty principle.

particle, going to zero at certain points. Since the momentum p is definite, the location of the particle must be equally probable at all points in space. Thus we reject the attempted wavefunction as inconsistent with the uncertainty principle.

The simplest wavefunction that is consistent with the wave-particle duality picture is $\psi_p(x) = Ae^{ipx/\hbar}$. The complex exponential respects the wave-nature of the particle by providing a periodic variation in x, yet it never goes to zero. The probability (density) is $|\psi_p(x)|^2 = |A|^2$, equal at all x. Thus, complex numbers are *inevitable* in the construction of the wavefunction representing a particle.



By linear superposition of complex exponentials, we can create 'particle' like or 'wave' like states as desired for the problem.

$\psi(x)$
x_0
$\psi(x) = \sum_p A_p e^{i\frac{px}{\hbar}}$
x
Δx
$||$
$A_1 \cdot e^{i\frac{p_1 x}{\hbar}}$
λ_1
x
$A_2 \cdot (e^{i p_2 x/\hbar} +$
λ_2
x
$A_3 \cdot (e^{i p_3 x/\hbar} +$
λ_3
x
$\Rightarrow$
$|A_{p_1}|^2$
$|A_{p_2}|^2$
p_1 p_2 p_3 p_4 p_5 p_6 $\dots$ p
Δp

The best we can do to locate a "particle" ⇒ a 'wave packet'.

$\Rightarrow$ $\boxed{\psi(x) = \sum_p A_p e^{i\frac{px}{\hbar}}}$ is an allowed "wavefunction".

- Drawing on Fourier series, we realize that we can create any wavefunction shape to capture the correct physics of the problem. Note the corresponding reciprocal space weight distribution.

Fig. 3.11: The superposition principle allows us to create wavefunctions that can represent as 'wave-like' or as 'particle-like' states we want. Wave-like states have large Δx and small Δp, and particle-like states have small Δx and large Δp. All the while, they satisfy the uncertainty principle $\Delta x \Delta p \geq \hbar/2$.

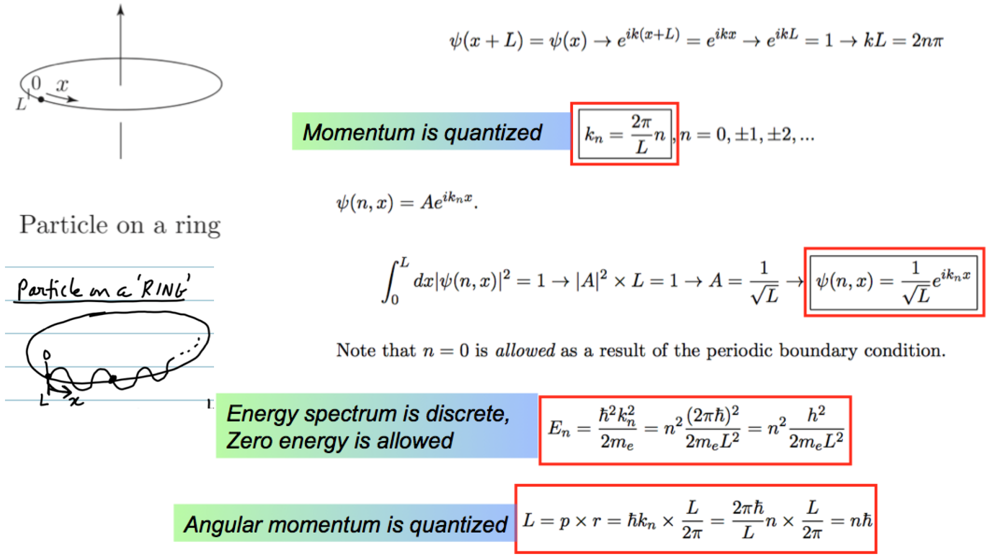
3.4 Operators

Every physical observable in quantum mechanics is represented by an operator. When the operator 'acts' on the wavefunction of the particle, it extracts the value of the observable. For example, the momentum operator is $\hat{p} = -i\hbar\partial/\partial x$, and for states of definite momentum $\hat{p}\psi_p(x) = (\hbar k)\psi_p(x)$. We note that $(x\hat{p} - \hat{p}x)f(x) = i\hbar f(x)$ for *any* function $f(x)$. The presence of the function in this equation is superfluous, and thus one gets the identity

$$x\hat{p} - \hat{p}x = [x, \hat{p}] = i\hbar. \tag{3.1}$$

The square brackets define a commutation relation. The space and momentum operators do not commute. In classical mechanics, $[x, p] = 0$. Quantum mechanics elevates the 'status' of x and p to those of mathematical operators, preventing them from commuting. This is referred to as the 'first quantization' from classical to quantum mechanics. In this scheme, the dynamical variables (x, p) that were scalars in classical mechanics are promoted to operators, and the wavefunction ψ is a scalar. If the number of particles is not conserved, then one needs to go one step further, and elevate the status of the wavefunction $\psi \to \hat{\psi}$ too, which is called second quantization.

3.5 States of definite momentum and location



$$\psi(x+L)=\psi(x)\rightarrow e^{ik(x+L)}=e^{ikx}\rightarrow e^{ikL}=1\rightarrow kL=2n\pi$$

Momentum is quantized $\boxed{k_n=\frac{2\pi}{L}n}, n=0,\pm1,\pm2,...$

$\psi(n,x)=Ae^{ik_nx}.$

$$\int_0^L dx|\psi(n,x)|^2=1\rightarrow|A|^2\times L=1\rightarrow A=\frac{1}{\sqrt{L}}\rightarrow\boxed{\psi(n,x)=\frac{1}{\sqrt{L}}e^{ik_nx}}$$

Note that $n=0$ is *allowed* as a result of the periodic boundary condition.

Energy spectrum is discrete, Zero energy is allowed $\boxed{E_n=\frac{\hbar^2k_n^2}{2m_e}=n^2\frac{(2\pi\hbar)^2}{2m_eL^2}=n^2\frac{h^2}{2m_eL^2}}$

Angular momentum is quantized $\boxed{L=p\times r=\hbar k_n\times\frac{L}{2\pi}=\frac{2\pi\hbar}{L}n\times\frac{L}{2\pi}=n\hbar}$

Fig. 3.12: Quantum mechanics of the particle on a ring.

The wavefunction $\psi_p(x)=Ae^{ipx/\hbar}$ is a state of definite momentum since it is an eigenstate of the momentum operator $\hat{p}\psi_p(x)=p\psi_p(x)$. One may demand the location of the particle to be limited to a finite length L. This may be achieved by putting an electron on a ring of circumference L, which yields upon normalization $A=1/\sqrt{L}$. In that case, the wavefunction must satisfy the relation $\psi_p(x+L)=\psi_p(x)$ to be single-valued. This leads to $e^{ikL}=1=e^{i2\pi\times n}$, and $k_n=n\times(2\pi/L)$. Here $n=0,\pm1,\pm2,...$ The linear momentum of the electron is then *quantized*, allowing only discrete values. Since $L=2\pi R$ where R is the radius of the ring, $k_nL=2\pi n\rightarrow pR=n\hbar$, showing angular momentum is quantized to $0,\pm\hbar,\pm2\hbar,....$ This indeed is the *quantum* of quantum mechanics! One may then index the wavefunctions of definite linear momentum by writing $\psi_n(x)$. Expressing states of definite momentum in terms of states of definite location similarly yields

$$\psi_n(x)=\frac{1}{\sqrt{L}}e^{ik_nx} \tag{3.2}$$

The set of wave functions $[...\psi_{-2}(x),\psi_{-1}(x),\psi_0(x),\psi_1(x),\psi_2(x),...]=[\psi_n(x)]$ are special. We note that $\int_0^L dx\psi_m^\star(x)\psi_n(x)=\delta_{nm}$, i.e., the functions are orthogonal. *Any* general wavefunction representing the particle $\psi(x)$ can be expressed as a linear combination of this set. This is the principle of superposition, and a basic mathematical result from Fourier theory. Thus the quantum mechanical state of a particle may be represented as $\psi(x)=\sum_n A_n\psi_n(x)$. Clearly, $A_n=\int dx\psi_n^\star(x)\psi(x)$. Every wavefunction constructed in this fashion represents a permitted state of the particle, as long as $\sum_n|A_n|^2=1$.

It is useful here to draw an analogy to the decomposition of a vector into specific coordinates. The 'hybrid' state function $\psi(x)$ is pictured as a vector $|\psi\rangle$ in an abstract space. The definite momentum wavefunctions $\psi_n(x)$ are pictured as the 'coordinate' vectors $|n\rangle$ in that space of vectors. This set of vectors is called the basis. Since there are an infinite set of integers $n=0,\pm1,\pm2,...$, the vector space is infinite dimensional. It is called the Hilbert space. One may then consider the coefficients A_n as the length of the projections of the state on the basis states. The abstract picture allows great economy of expression by writing $|\psi\rangle=\sum_n A_n|n\rangle$. The orthogonality of the basis states is $\langle m|n\rangle=\delta_{mn}$, and thus $A_n=\langle n|\psi\rangle$. Then it is evident that $|\psi\rangle=\sum_n\langle n|\psi\rangle|n\rangle=\sum_n|n\rangle\langle n|\psi\rangle$, and $\sum_n|n\rangle\langle n|=1$.

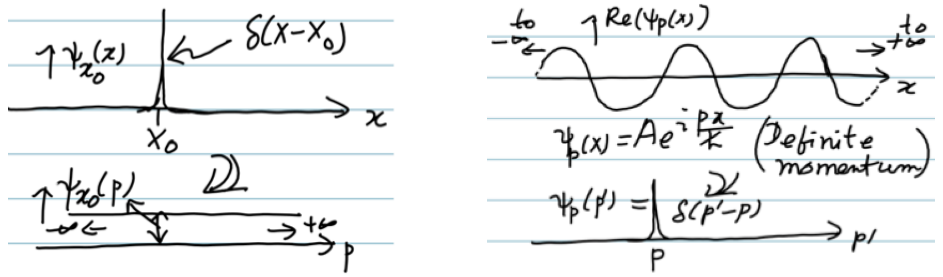
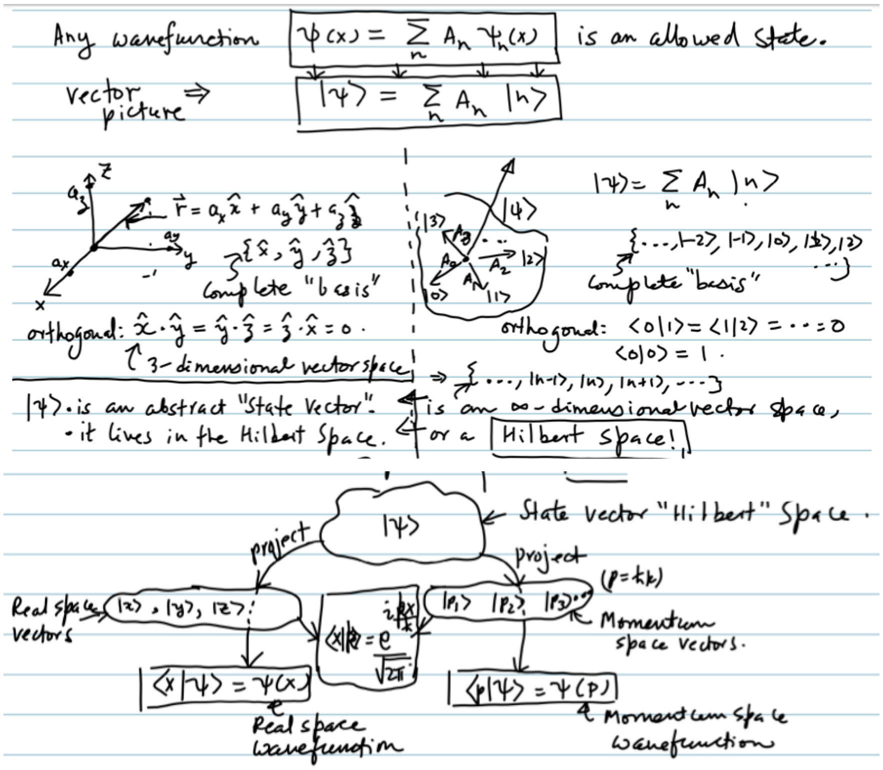


Fig. 3.13: States of definite location and states of definite momentum.



Any wavefunction
$\psi(x) = \sum_n A_n \psi_n(x)$
is an allowed state.
vector picture $\Rightarrow$
$|\psi\rangle = \sum_n A_n |n\rangle$
z
a_z
a_y
y
a_x
x
$\vec{r} = a_x\hat{x} + a_y\hat{y} + a_z\hat{z}$
$\{\hat{x}, \hat{y}, \hat{z}\}$
Complete "basis"
orthogonal: $\hat{x}\cdot\hat{y} = \hat{y}\cdot\hat{z} = \hat{z}\cdot\hat{x} = 0$.
3-dimensional vector space
$|3\rangle$
A_3
$|\psi\rangle$
A_0
A_2
$|2\rangle$
A_1
$|0\rangle$
$|1\rangle$
$|\psi\rangle = \sum_n A_n |n\rangle$
$\{\ldots, |-2\rangle, |-1\rangle, |0\rangle, |1\rangle, |2\rangle \ldots\}$
Complete "basis"
orthogonal: $\langle 0|1\rangle = \langle 1|2\rangle = \cdots = 0$
$\langle 0|0\rangle = 1$.
$\Rightarrow \{\ldots, |n-1\rangle, |n\rangle, |n+1\rangle, \ldots\}$
$|\psi\rangle$. is an abstract "State Vector".
• it lives in the Hilbert Space.
is an ∞-dimensional vector space,
or a Hilbert space!
$|\psi\rangle$
State vector "Hilbert" Space.
project
project
$(p=\hbar k)$
Real space vectors
$|x\rangle, |y\rangle, |z\rangle$
$\langle x|p\rangle = \frac{e^{i\frac{px}{\hbar}}}{\sqrt{2\pi}}$
$|p_1\rangle$ $|p_2\rangle$ $|p_3\rangle \cdots$
Momentum space vectors.
$\langle x|\psi\rangle = \psi(x)$
Real space wavefunction
$\langle p|\psi\rangle = \psi(p)$
Momentum space wavefunction

Fig. 3.14: Vector spaces for quantum states: we can use results of linear algebra for quantum mechanics problems.

A vector may be decomposed in various basis coordinates. For example, a vector in 3-d real space may be decomposed into cartesian, spherical, or cylindrical coordinate systems. Similarly, the choice of basis states of definite momentum is not unique. The wavefunctions for states of definite location are those functions that satisfy $x\psi_{x_0}(x) = x_0\psi_{x_0}(x)$, which lets us identify $\psi_{x_0}(x) = \delta(x - x_0)$. Here $\delta(...)$ is the Dirac-delta function, sharply peaked at $x = x_0$. It is instructive to expand the states of definite location in the basis of the states of definite momentum. From the uncertainty relation, we expect a state of definite location to contain many momenta. The expansion yields $A_n = \int_{-\infty}^{+\infty} dk/(2\pi/L) \times (e^{ik_nx}/\sqrt{L})\delta(x - x_0) = e^{ik_nx_0}/\sqrt{L}$, whereby $|A_n|^2 = 1/L$. Thus, the state of definite location x_0 is constructed of an infinite number of states of definite momentum $n = 0, \pm1, \pm2, ...$, each with equal probability $1/L$.

3.6 States of definite energy: The Schrodinger equation

States of definite energy $\psi_E(x)$ are special. Unlike the states of definite momentum or definite location, we cannot write down their general wavefunction without more information. That is because the energy of a particle depends on its potential and kinetic components. In classical mechanics, the total energy is $p^2/2m + V(x)$, i.e., split between kinetic and potential energy components. Once x & p are known for a classical particle, the energy is completely defined, meaning one does not need to ask another question. However, since x and p cannot be simultaneously defined for a quantum-mechanical particle with arbitrary accuracy, the energy must be obtained through operations performed on the wavefunction.

Schrodinger provided the recipe, and the equation is thus identified with his name. The Schrodinger equation is

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)\right]\psi_E(x) = E\psi_E(x). \tag{3.3}$$

Fig. 3.15: Erwin Schrodinger introduced 'wave equation' for quantum mechanics. Nobel prize in 1933.

The solution of this eigenvalue equation for a $V(x)$ identifies the special wavefunctions $\psi_E(x)$. These wavefunctions represent states of definite energy. How did we ascertain the accuracy of the Schrodinger equation? The answer is through experiments. A major unresolved problem at the time was explaining the discrete spectral lines emitted from excited hydrogen atoms. Neils Bohr had a heuristic model to explain the spectral lines that lacked mathematical rigor. The triumph of Schrodinger equation was in explaining the precise spectral lines. An electron orbiting a proton in a hydrogen atom sees a potential $V(r) = -q^2/4\pi\epsilon_0 r$. Schrodinger solved this equation (with help from a mathematician), and obtained energy eigenvalues $E_n = -13.6/n^2$ eV. Thus Bohr's semi-qualitative model was given a rigid mathematical basis by Schrodinger's equation. The equation also laid down the recipe for solving similar problems in most other situations we encounter. Just as the case for states of definite energy or definite location, one may expand *any* state of a quantum particle in terms of the states of definite energy $\psi(x) = \sum_E A_E\psi_E(x)$, or equivalently $|\psi\rangle = \sum_E A_E|E\rangle$

So why do states of definite energy occupy a special position in applied quantum mechanics? That becomes clear if we consider the time-dependent Schrodinger equation.

Fig. 3.16: Niels Bohr was the original architect and 'conscience keeper' of quantum mechanics, and an intellectual leader who influenced an entire generation. Nobel prize in 1922.

3.7 Time-dependent Schrodinger equation

Newton's law $F = dp/dt$ provides the prescription for determining the future (x', p') of a particle given its present (x, p). Schrodinger provided the quantum-mechanical equivalent, through the time-dependent equation

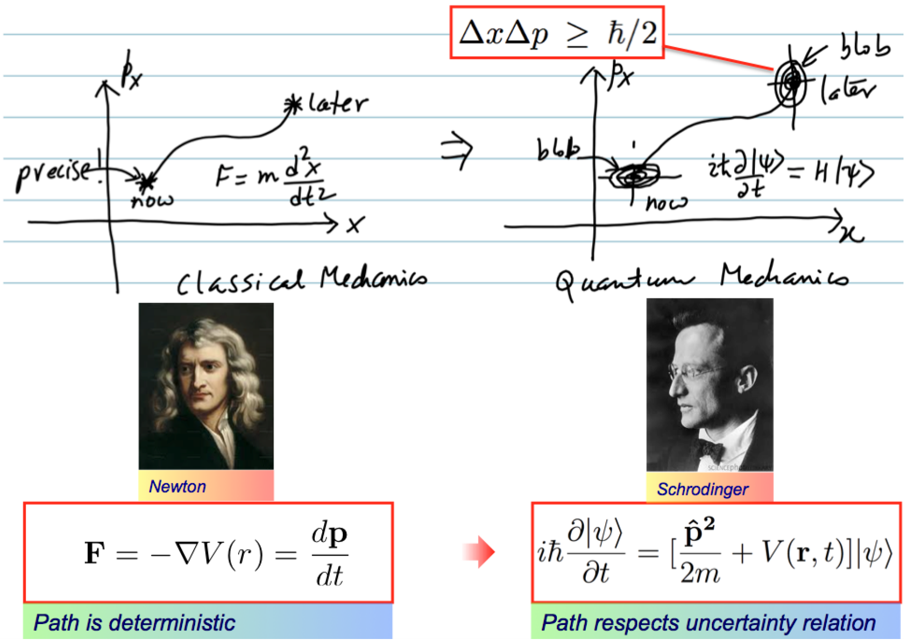

$\Delta x \Delta p \geq \hbar/2$
p_x
later
precise
now
$F = m \frac{d^2x}{dt^2}$
x
Classical Mechanics
$\Rightarrow$
blob
p_x
blob
later
blob
now
$i\hbar \frac{\partial |\psi\rangle}{\partial t} = H|\psi\rangle$
x
Quantum Mechanics
Newton
Schrodinger
$\mathbf{F} = -\nabla V(r) = \frac{d\mathbf{p}}{dt}$
Path is deterministic
$i\hbar \frac{\partial |\psi\rangle}{\partial t} = [\frac{\hat{\mathbf{p}}^2}{2m} + V(\mathbf{r}, t)]|\psi\rangle$
Path respects uncertainty relation

Fig. 3.17: The dynamics of quantum states is governed by the time-dependent Schrodinger equation. Note that it looks like a hybrid of the classical energy and a wave equation, which is how it must be to account for the wave-particle duality.

$$i\hbar \frac{\partial \Psi(x,t)}{\partial t} = \underbrace{[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)]}_{\hat{H}}\Psi(x,t). \tag{3.4}$$

To track the time-evolution of quantum states, one must solve this equation and obtain the composite space-time wavefunction $\Psi(x,t)$. Then physical observables can be obtained by operating upon the wavefunction by the suitable operators. Let's look at a particular set of solution wavefunctions which allow the separation of the time and space variables, of the form $\Psi(x,t) = \chi(t)\psi(x)$. Inserting it back into the time-dependent Schrodinger equation and rearranging, we obtain

$$i\hbar \frac{\dot{\chi(t)}}{\chi(t)} = \frac{\hat{H}\psi(x)}{\psi(x)} = E. \tag{3.5}$$

Note that since the left side does not depend on space, and the right side does not depend on time, both the fractions must be a constant. The constant is called E, and clearly has dimensions of energy in Joules. The right half of the equation lets us identify that $\hat{H}\psi_E(x) = E\psi_E(x)$ are states of definite energy. Then the left side dictates that the time dependence of these states is described by $\chi(t) = \chi(0)e^{-iEt/\hbar}$. Thus the particular set of solutions

$$\Psi_E(x,t) = \psi_E(x)e^{-i\frac{E}{\hbar}t} \tag{3.6}$$

now define the time evolution of the states. Here $\psi_E(x)$ are states of definite energy, as obtained by solving the time-independent Schrodinger equation.

3.8 Stationary states and time evolution

We note that $|\Psi_E(x,t)|^2 = |\psi_E(x)|^2$, that is, the state $\Psi_E(x,t)$ does *not* change with time. That means that a particle prepared in a state of definite energy will stay in that energy if there are no perturbations. Its wavefunction does evolve as $\exp(-iEt/\hbar)$, but this evolution is 'unitary' since its absolute value is unity. Notice the analogy with Newton's first law, which states that a particle at rest or moving with constant velocity will continue to do so unless acted upon by a force. The states of definite energy are therefore special since they do not evolve with time unless perturbed, and are called 'stationary states'. Thus the expansion may be written as

$$\Psi(x,t) = \sum_E A_E \Psi_E(x,t) = \sum_E A_E \psi_E(x) e^{-i\frac{E}{\hbar}t}. \tag{3.7}$$

The states of definite energy form a convenient and often-used basis for expansion of general states of a particle. That is because they are stationary states - it is simpler if the basis states are fixed.

Consider a simple case where a hybrid state $\Psi(x,t)$ is prepared with components in two states $|E_1\rangle$ and $|E_2\rangle$. Then, the expansion is $\Psi(x,t) = A_{E_1}\psi_{E_1}(x)e^{-iE_1 t/\hbar} + A_{E_2}\psi_{E_2}(x)e^{-iE_2 t/\hbar}$. The probability density of this state then is, for real A's

$$|\Psi(x,t)|^2 = |A_{E_1}|^2|\psi_{E_1}(x)|^2 + |A_{E_2}|^2|\psi_{E_2}(x)|^2 + A_{E_1}A_{E_2}\psi_{E_1}(x)\psi_{E_2}(x)\cos\left(\frac{E_1 - E_2}{\hbar}t\right), \tag{3.8}$$

which *does* oscillate with time with frequency $\omega_{12} = (E_1 - E_2)/\hbar$. Such two-level systems are being currently explored for making quantum-bits or qubits for a form of analog computation called quantum-computation.

All transport and optical phenomena involve time evolution. So most of the time in semiconductor physics we we are working with the solutions of the time-dependent Schrodinger equation. The states of definite energy as a function of momentum $E(k)$ that form the *energy bandstructure* of the solid thus provide a most convenient basis for the analysis of electronic and optical phenomena of semiconductors.

The time evolution of the expectation value of an operator is given by Ehrenfrest's theorem

$$\frac{d\langle \hat{A}\rangle}{dt} = -\frac{i}{\hbar}\langle[\hat{A},\hat{H}]\rangle, \tag{3.9}$$

where the operator itself is time-independent. By using $\hat{A} = \hat{p}$ and $\hat{H} = \hat{p}^2/2m + V(x)$, Ehrenfrest's theorem directly leads to Newton's 2nd law. It forms the starting point for the density-matrix formulation of the time-evolution of quantum states.

3.9 Quantum Current

In semiconductor devices, we will be deeply concerned with the flow of currents. A current is a measure of the flow of objects from one point in space to another. The flow of electric charges constitutes an electric current, leading to the notion of electrical conductivity. In this chapter we develop the recipe to understand current flow from a quantum-mechanical viewpoint. Since the physical state of particles in quantum mechanics is represented by its wavefunction $\Psi(x,t)$, the current must be obtained from the wavefunction.

Since $|\Psi(x,t)|^2 = \Psi^\star\Psi$ is the probability density, let's examine how it changes with time. We obtain

$$\frac{\partial|\Psi(x,t)|^2}{\partial t} = \Psi^\star\frac{\partial\Psi}{\partial t} + \frac{\partial\Psi^\star}{\partial t}\Psi, \tag{3.10}$$

where we use the time-dependent Schrodinger equation $i\hbar\partial\Psi/\partial t = (\hat{p}^2/2m + V)\Psi$ and its complex conjugate $-i\hbar\partial\Psi^\star/\partial t = (\hat{p}^2/2m + V)\Psi^\star$ to obtain

$$\frac{\partial|\Psi(x,t)|^2}{\partial t} = \Psi^\star\frac{(\hat{p}^2/2m + V)\Psi}{i\hbar} + \Psi\frac{(\hat{p}^2/2m + V)\Psi^\star}{-i\hbar}, \tag{3.11}$$

which simplifies to

$$\frac{\partial|\Psi(x,t)|^2}{\partial t} = \frac{1}{2mi\hbar}(\Psi^\star\hat{p}^2\Psi - \Psi\hat{p}^2\Psi^\star). \tag{3.12}$$

Since $\hat{p} = -i\hbar\nabla_{\mathbf{r}}$, we recognize the resulting equation

$$\frac{\partial|\Psi(x,t)|^2}{\partial t} = -\nabla_{\mathbf{r}}\cdot\big[\frac{1}{2m}(\Psi^\star\hat{p}\Psi - \Psi\hat{p}\Psi^\star)\big] \tag{3.13}$$

as the familiar 'continuity' equation in disguise. A continuity equation is of the form $\partial\rho/\partial t = -\nabla_{\mathbf{r}}\cdot\mathbf{j}$, where ρ is the particle 'density' and $\mathbf{j}$ is the current density. This is illustrated in Figure 3.18.

We read off the quantum mechanical current density as

$$\boxed{\mathbf{j} = \frac{1}{2m}(\Psi^\star\hat{\mathbf{p}}\Psi - \Psi\hat{\mathbf{p}}\Psi^\star).} \tag{3.14}$$

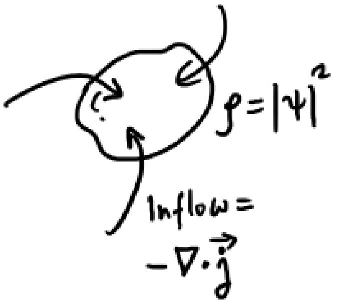


Fig. 3.18: Current continuity $\frac{\partial\rho}{\partial t} = -\nabla\cdot\mathbf{j}$.

This equation provides us the required recipe for calculating the probability density flow, or current flow directly from the quantum mechanical wavefunctions of states. We make a few observations. If Ψ is real, $\mathbf{j} = 0$. Since Ψ has dimension of $1/\sqrt{Vol}$, the dimension of $\mathbf{j}$ is per unit area per second. For 3D, volume is in m^3 and $\mathbf{j}$ is then in $1/(\mathrm{m}^2\cdot\ \mathrm{s})$. For 2D $\mathbf{j}$ is in $1/(\mathrm{m}\cdot\mathrm{s})$, and it is simply 1/s for 1D. We will use this concept of currents in greater detail in later chapters, and generalize it to charge, heat, or spin currents.

We also note that

$$\frac{d}{dt}(\int_{space} d^3r|\Psi|^2) = -\int_{space} d^3r\nabla\cdot\mathbf{j} = -\oint\mathbf{j}\cdot d\mathbf{S} = 0. \tag{3.15}$$

The conversion of the integral from volume to a closed surface uses Gauss' theorem. The value of the integral is zero because Ψ and consequently $\mathbf{j}$ goes to zero at infinity, and the equality must hold for all space. This equation is a statement of the indestructibility of the particle, which follows from $\int_{space} d^3r|\Psi|^2 = 1$. If the number of particles is *not* conserved, then one needs to add recombination ('annihilation') and generation ('creation') terms to the continuity equation. It then looks as $\partial\rho/\partial t = -\nabla\cdot\mathbf{j} + (G - R)$ where R and G are recombination and generation rates.

We also note that in the presence of a magnetic field $\mathbf{B} = \nabla\times\mathbf{A}$, the quantum-mechanical momentum operator $\hat{\mathbf{p}} \to \hat{\mathbf{p}} + q\mathbf{A}$ where q is the magnitude of the electron charge. This leads to an additional term in the expression of the current density

$$\mathbf{j} = \frac{1}{2m}(\Psi^\star\hat{\mathbf{p}}\Psi - \Psi\hat{\mathbf{p}}\Psi^\star) + \frac{q\mathbf{A}}{m}\Psi^\star\Psi. \tag{3.16}$$

The additional term depending on the magnetic vector potential $\mathbf{A}$ is useful to explain current flow in magnetic materials, magnetotransport properties, and superconductivity.

If we want to determine the electric charge current, we realize that the current flux is actually of electrons that have wavefunction Ψ for which we have calculated the probability current flux $\mathbf{j}$. The charge q is dragged along by the electron. So to account for the flow of *charge*, the current density is simply $\mathbf{J} = q\mathbf{j}$, where q is the charge (in Coulombs) of the charge particle. If these charge particles are electrons, $q = 1.6\times 10^{-19}$ C and free mass

$m_e = 9.1 \times 10^{-31}$ kg. In the absence of a magnetic field, the electric current density is then given by

$$\boxed{\mathbf{J} = \frac{q}{2m_e}(\Psi^\star \hat{\mathbf{p}} \Psi - \Psi \hat{\mathbf{p}} \Psi^\star),} \tag{3.17}$$

which is now in A/m^2 for 3D, A/m for 2D, and A for 1D current flow, where A=C/s is the unit of current in Amperes. The current density is expressed in terms of the electron wavefunctions. We wish to make the expression more 'usable'.

Consider free electrons in 1D with periodic boundary conditions between $x = (0, L)$. The wavefunction for a state $|k\rangle$ of definite energy $E(k)$ is $\Psi_E(x,t) = (1/\sqrt{L})e^{ikx}e^{-iE(k)t/\hbar}$. In the QM expression for current, the time evolution portion is not affected by the momentum operator, and therefore factors to 1. It is another illustration of the virtues of working with states of definite energy. The current carried by state $|k\rangle$ is then obtained as $J(k) = I(k) = q\hbar k/m_e L$. The current density and current are the same in 1D. The current $I(k) = q\hbar k/m_e L = qv(k)/L$ connects to the classical notion of current carried by a particle with velocity $v(k) = \hbar k/m_e$ traversing a distance L. Another way to picture the same current is to split it as $I = q \times v(k) \times n$, where $n = 1/L$ is the 'volume density' of particles.

So we can find the current flow due to each allowed $k-$state for any quantum particle. Now let $f(k)$ be an occupation function that determines whether that $k-$state is occupied by a particle or not, and if it is, how many particles are sitting in it. To find the occupation function $f(k)$, we stumble upon one of the deepest mysteries that that was unearthed by quantum mechanics.

3.10 Fermions and Bosons

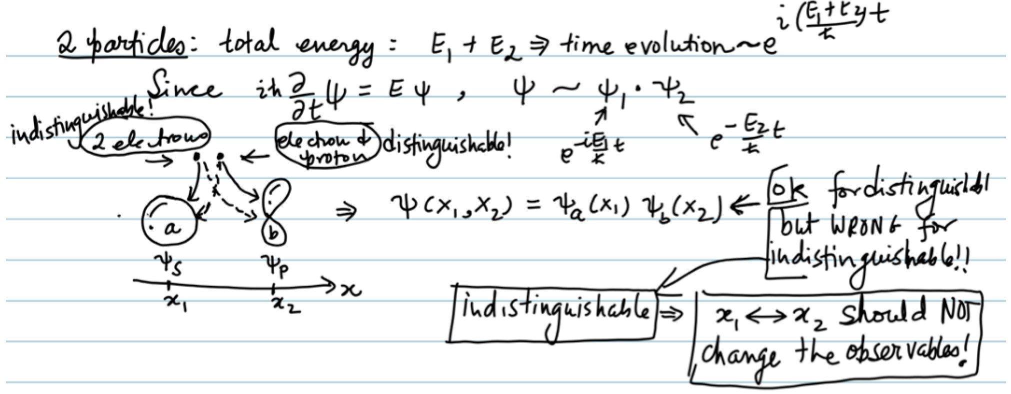

2 particles: total energy : $E_1 + E_2 \Rightarrow$ time evolution $\sim e^{i(\frac{E_1+E_2}{\hbar})t}$
Since $i\hbar\frac{\partial}{\partial t}\psi = E\psi$, $\psi \sim \psi_1 \cdot \psi_2$
$e^{-i\frac{E_1}{\hbar}t}$
$e^{-\frac{E_2}{\hbar}t}$
indistinguishable!
2 electrons
electron + proton
distinguishable!
a
b
ψ_s
ψ_p
x
x_1
x_2
$\Rightarrow \psi(x_1, x_2) = \psi_a(x_1)\,\psi_b(x_2)$
ok for distinguishabl
but WRONG for indistinguishable!!
indistinguishable $\Rightarrow$
$x_1 \leftrightarrow x_2$ should NOT change the observables!

Fig. 3.19: Indistinguishable particles suffer an identity crisis when we try constructing a wavefunction for more than one particle!

Consider two quantum states $|a\rangle$ and $|b\rangle$ with real-space wavefunctions $\psi_a(x)$ and $\psi_b(x)$. What is the many-particle wavefunction when *two* particles are put in the two states? Lets label the locations of the two particles as x_1 and x_2. If the two particles are *distinguishable*, such as an electron and a proton, then the composite wavefunction may be written as the product of the single-particle wavefunctions

$$\psi(x_1, x_2) = \psi_a(x_1)\psi_b(x_2). \tag{3.18}$$

But if the two particles are *indistinguishable*, such as two electrons, the wavefunction must satisfy further requirements. Specifically, if we swap the locations of the two electrons

$x_1 \leftrightarrow x_2$, the physical observables of the composite state must remain the same. This requirement dictates that the probability density must satisfy

$$P(x_2, x_1) = P(x_1, x_2) \rightarrow |\psi(x_2, x_1)|^2 = |\psi(x_1, x_2)|^2. \tag{3.19}$$

The original product wavefunction does not satisfy this requirement. It cannot represent indistinguishable particles. A symmetrized form, however does the job:

$$\psi(x_1, x_2) = \psi_a(x_1)\psi_b(x_2) + \psi_a(x_2)\psi_b(x_1) \tag{3.20}$$

because

$$\psi(x_2, x_1) = +\psi(x_1, x_2) \tag{3.21}$$

and the probability density does not change upon swapping. We also note that *both* particles may be in the same x since

$$\psi(x_1, x_1) = +\psi(x_1, x_1) \tag{3.22}$$

is OK. Particles in nature that choose the '+' sign are bosons. Multiple bosons can occupy the sate quantum state.

Fig. 3.20: Wolfgang Pauli discovered the exclusion principle for which he won the Nobel prize in 1945. Introduced matrices for electron spin. Humorously referred to as the imaginary part of another notable physicist - Wolfgang Paul.

The anti-symmetrized form

$$\psi(x_1, x_2) = \psi_a(x_1)\psi_b(x_2) - \psi_a(x_2)\psi_b(x_1) \tag{3.23}$$

leads to

$$\psi(x_2, x_1) = -\psi(x_1, x_2), \tag{3.24}$$

which is also permitted, since the probability density remains unaltered by the negative sign upon swapping the particles. Particles that choose the '-' sign are fermions. However, an attempt to put both fermions in the same location leads to

$$\psi(x_1, x_1) = -\psi(x_1, x_1) \rightarrow \psi(x_1, x_1) = 0. \tag{3.25}$$

This is the Pauli exclusion principle. It states the simple result that two identical fermions (e.g. electrons) cannot be in the same quantum state. It is responsible for all chemical behavior of matter and the existence of the periodic table of elements.

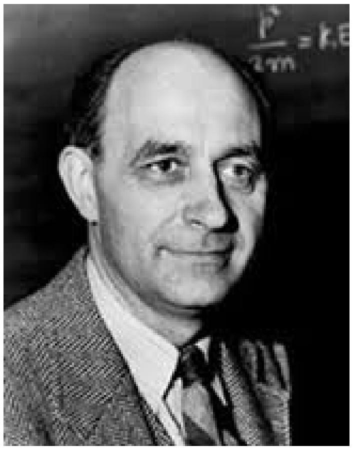


Fig. 3.21: Enrico Fermi made significant contributions to virtually all fields of physics. Nobel prize in 1938. The frequency of occurrence of his name in this book is proof enough of his influence.

In the presence of large number of electrons, the Pauli-exclusion principle leads to an occupation probability of quantum states. The result was first derived by Dirac, and is called the Fermi-Dirac relation

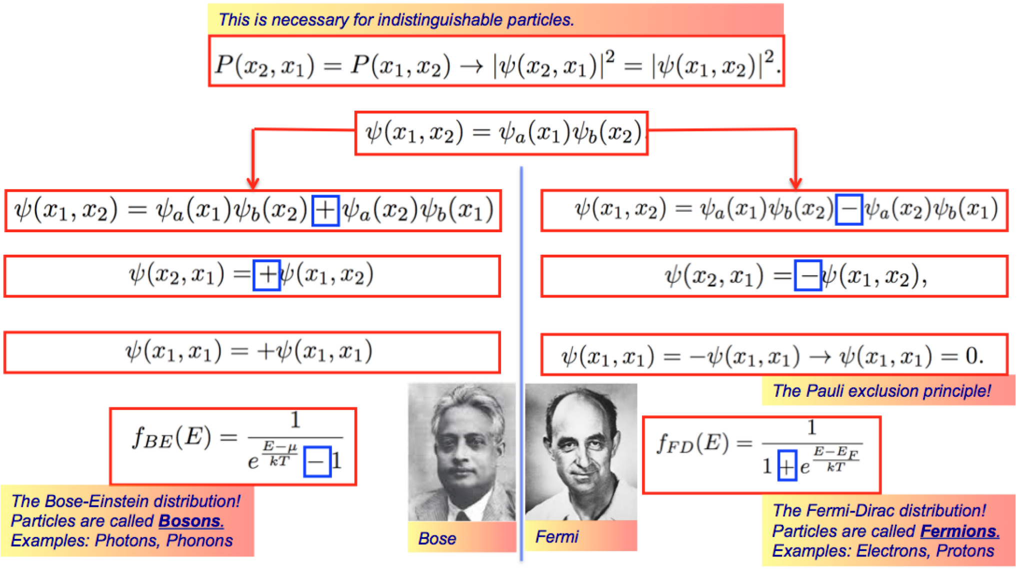
$$f_{FD}(E) = \frac{1}{1 + e^{\frac{E-E_F}{kT}}}, \tag{3.26}$$

where E_F is the Fermi-energy, k the Boltzmann constant, and T the absolute temperature. Note that the value cannot exceed 1.

The equivalent statistical result for bosons is

$$f_{BE}(E) = \frac{1}{e^{\frac{E-\mu}{kT}} - 1}, \tag{3.27}$$

where μ is the chemical potential. The Bose-Einstein distribution allows values larger than 1. Dramatic effects such as the Bose-Einstein condensation (BEC), lasers, and the existence of superconductivity occurs when many bosons can co-exist in the same state. The bosons can be composite particles, for example Cooper-pairs in superconductors that are electron-phonon-electron quasiparticles where electrons are 'glued' together by phonons.



This is necessary for indistinguishable particles.

$$P(x_2,x_1)=P(x_1,x_2)\to|\psi(x_2,x_1)|^2=|\psi(x_1,x_2)|^2.$$

$$\psi(x_1,x_2)=\psi_a(x_1)\psi_b(x_2)$$

$$\psi(x_1,x_2)=\psi_a(x_1)\psi_b(x_2)\boxed{+}\psi_a(x_2)\psi_b(x_1)$$

$$\psi(x_2,x_1)=\boxed{+}\psi(x_1,x_2)$$

$$\psi(x_1,x_1)=+\psi(x_1,x_1)$$

$$f_{BE}(E)=\frac{1}{e^{\frac{E-\mu}{kT}}\boxed{-}1}$$

The Bose-Einstein distribution!
*Particles are called **Bosons.***
Examples: Photons, Phonons

$$\psi(x_1,x_2)=\psi_a(x_1)\psi_b(x_2)\boxed{-}\psi_a(x_2)\psi_b(x_1)$$

$$\psi(x_2,x_1)=\boxed{-}\psi(x_1,x_2),$$

$$\psi(x_1,x_1)=-\psi(x_1,x_1)\to\psi(x_1,x_1)=0.$$

The Pauli exclusion principle!

$$f_{FD}(E)=\frac{1}{1\boxed{+}e^{\frac{E-E_F}{kT}}}$$

The Fermi-Dirac distribution!
*Particles are called **Fermions.***
Examples: Electrons, Protons

Fig. 3.23: Indistinguishable particles can be of two types: Bosons, or Fermions. They have very different properties!

3.11 Spin, and the Spin-Statistics Theorem

In addition to linear momentum $\mathbf{p}=\hbar\mathbf{k}$ and angular momentum $\mathbf{L}=\mathbf{r}\times\mathbf{p}$, electrons also possess an extra bit of spin angular momentum. In semiconductors, electron spin plays an important role in the electronic band structure. The net angular momentum of electron states is obtained by adding the various components of the angular momenta.

The exclusion principle is central to the spin-statistics theorem from relativistic quantum field-theory. It states that bosonic particles have integer spins, and fermonic particles have half-integer spins. That means bosons have spins $S=0,\pm\hbar,\pm2\hbar,...$, and fermions have spins $S=\pm\hbar/2,\pm3\hbar/2,....$ Electrons have spin $\pm\hbar/2$.

The fundamental dichotomy of particles in nature has received increasing attention the last three decades. Quasi-particle states have been observed (for example in the fractional quantum Hall effect) that behave neither like fermions nor bosons. Swapping the single-particle states for such quasi-particles leads to the accumulation of a phase factor:

$$\psi(x_2,x_1)=e^{i\phi}\psi(x_1,x_2).\tag{3.28}$$

These particles evidently satisfy the indistinguishability criteria, but accumulate a(ny) phase, leading to their name *anyons*. Anyon states can exhibit a richer range of statistics than fermions and bosons. For anyons, commuting (or Abelian) statistics has similarity to fermions and bosons, but non-commuting (or non-Abelian) statistics does not have such an analog. Non-Abelian anyons are of current interest due to their proposed usage in topological quantum computation.

Fig. 3.22: Satyendra Nath Bose, who discovered the statistics for photons. Particles that follow such statistics are now called Bosons.

3.12 The Dirac equation and the birth of particles

Dirac was not comfortable with Schrodinger's equation since it was not consistent with relativity, and did not predict spin of electrons. He was able to reformulate the quantum-mechanics of electrons from Schrodinger's equation

$$i\hbar\frac{\partial|\psi\rangle}{\partial t} = [\frac{\hat{\mathbf{p}}^{\mathbf{2}}}{2m} + V(\mathbf{r},t)]|\psi\rangle \tag{3.29}$$

to the Dirac equation

$$i\hbar\frac{\partial|\psi\rangle}{\partial t} = [c\alpha\cdot\hat{\mathbf{p}} + \beta mc^2 + V(\mathbf{r},t)]|\psi\rangle \tag{3.30}$$

where c is the speed of light, and $\hat{\alpha},\beta$ are matrices. Before Dirac, the concept of a 'particle' was not very clear. Dirac's assertion was to the effect: 'a particle is the solution of my equation'. Dirac's equation described the electron energy spectrum with more accuracy than Schrodinger's equation, and accounted for spin naturally. It also predicted the existence of negative energy states, or anti-electrons. This was the first prediction of antimatter. A few years after the prediction, such particles were discovered in cloud chambers by Carl Anderson; these particles are called positrons. Electrons and positrons annihilate each other, emitting light of energy $\hbar\omega = 2m_0c^2$.

The philosophy of Dirac that 'particles are solutions to equations' gave rise to the prediction of a number of new particles that have since been observed such as quarks, gluons, Higgs boson, etc... Majorana fermions fall under the category of predicted exotic particles, and there is intense interest in realizing such exotic states in matter for topological quantum computation. What was exotic yesterday will become commonplace tomorrow, so keep track of those 'particles'!

Fig. 3.24: Paul Dirac unified quantum theory with special relativity and discovered an equation that predicted the spin of electron. Shared the physics Nobel prize in 1933 with Schrodinger.

Chapter Summary

The five basic postulates of quantum mechanics are:

(1) The state of any physical system at a given time t is completely represented by a state vector $|\Psi\rangle = |\Psi(\mathbf{r},t)\rangle$.

(2) For a physically observervable quantity A there is an operator $\hat{\mathbf{A}}$. The eigenvalues of $\hat{\mathbf{A}}$ are the possible results of the measurements of A, that is, denoting the eigenvalues of $\hat{\mathbf{A}}$ by a,

$$\hat{\mathbf{A}}|a\rangle = a|a\rangle, \tag{3.31}$$

and the probability of a measurement of A yielding the value a at time t is $|\langle a|\Psi(t)\rangle|^2$. The a's, which are the results of possible measurements, must be real. This implies that $\hat{\mathbf{A}}$ must be a linear hermitian operator.

(3) A measurement of $|\Psi\rangle$ that leads to an eigenvalue a_i leads the quantum mechanical system to *collapse* into the eigenstate $|\Psi_i\rangle$, which is the eigenstate corresponding to the eigenvalue a_i. So a measurement affects the state of the quantum system.

(4) There exists a hermitian operator $\hat{\mathbf{H}}$ such that

$$i\hbar\frac{\partial|\Psi(\mathbf{r},t)\rangle}{\partial t} = \hat{\mathbf{H}}|\Psi(\mathbf{r},t)\rangle. \tag{3.32}$$

(5) Two classical dynamical variables a, b, which are conjugate in the Hamiltonian sense, are represented by Schrodinger operators $\hat{\mathbf{A}}$,$\hat{\mathbf{B}}$, which obey

$$\hat{\mathbf{A}}_i\hat{\mathbf{B}}_j - \hat{\mathbf{B}}_j\hat{\mathbf{A}}_i = i\hbar\delta_{ij}. \tag{3.33}$$

Problems

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 4

Damned Lies, and Statistics

Contents

4.1 The physics of equilibrium 31
4.2 Partition Function for Quantum Systems . . 32
4.3 The Fermi-Dirac Distribution 33
4.4 The Bose-Einstein Distribution 34
4.5 Manifestations of the distribution functions 34
4.6 Meaning of equilibrium in semiconductor devices 37
Summary 40
Problems 41

In this chapter, we derive and discuss the Fermi-Dirac distribution function for fermions, and the Bose-Einstein distribution function for bosons. These functions provide us the statistical occupation number of quantum states for a system in thermodynamic equilibrium with a reservoir. The Fermi-Dirac distribution is central to finding the electron distribution over allowed energy or momentum values in various semiconductor devices. The Bose-Einstein distribution is central to finding the distribution of photons in the electromagnetic field, or phonons in semiconductor crystals. The two distributions together determine electron-phonon and electron-photon interactions. The importance of this chapter simply cannot be overemphasized! We discuss various properties of the distributions and limiting cases to gain familiarity. Then, we specifically map the concept of thermodynamic equilibrium to the fundamental semiconductor building blocks, such as the ohmic contact, Schottky contacts, the p-n junction, and a field-effect transistor (FET).

4.1 The physics of equilibrium

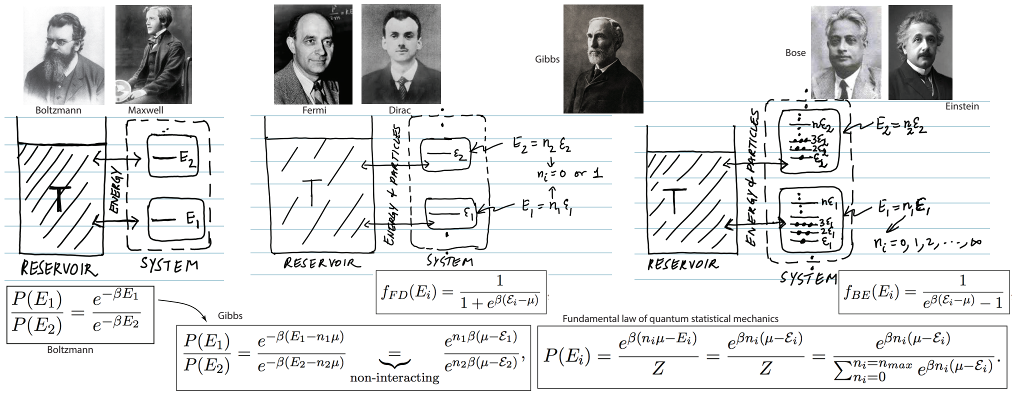

Boltzmann
Maxwell
Fermi
Dirac
Gibbs
Bose
Einstein
T
ENERGY
RESERVOIR
SYSTEM
E_2
E_1
T
ENERGY & PARTICLES
RESERVOIR
SYSTEM
$\mathcal{E}_2$
$\mathcal{E}_1$
$E_2 = n_2 \mathcal{E}_2$
$n_i = 0$ or 1
$E_1 = n_1 \mathcal{E}_1$
T
ENERGY & PARTICLES
RESERVOIR
SYSTEM
$n\mathcal{E}_2$
$3\mathcal{E}_2$
$2\mathcal{E}_2$
$\mathcal{E}_2$
$n\mathcal{E}_1$
$3\mathcal{E}_1$
$2\mathcal{E}_1$
$\mathcal{E}_1$
$E_2 = n_2 \mathcal{E}_2$
$E_1 = n_1 \mathcal{E}_1$
$n_i = 0, 1, 2, \cdots, \infty$
$\frac{P(E_1)}{P(E_2)} = \frac{e^{-\beta E_1}}{e^{-\beta E_2}}$
Boltzmann
$f_{FD}(E_i) = \frac{1}{1 + e^{\beta(\mathcal{E}_i - \mu)}}$
$f_{BE}(E_i) = \frac{1}{e^{\beta(\mathcal{E}_i - \mu)} - 1}$
Gibbs
$\frac{P(E_1)}{P(E_2)} = \frac{e^{-\beta(E_1 - n_1\mu)}}{e^{-\beta(E_2 - n_2\mu)}} \underbrace{=}_{\text{non-interacting}} \frac{e^{n_1\beta(\mu - \mathcal{E}_1)}}{e^{n_2\beta(\mu - \mathcal{E}_2)}},$
Fundamental law of quantum statistical mechanics
$P(E_i) = \frac{e^{\beta(n_i\mu - E_i)}}{Z} = \frac{e^{\beta n_i(\mu - \mathcal{E}_i)}}{Z} = \frac{e^{\beta n_i(\mu - \mathcal{E}_i)}}{\sum_{n_i=0}^{n_i=n_{max}} e^{\beta n_i(\mu - \mathcal{E}_i)}}.$

Fig. 4.1: Illustration of the processes of thermodynamic equilibrium for the Boltzmann distribution, and the Gibbs partition function.

We begin by drawing upon a fundamental result from quantum statistical mechanics[1]. The most well-known result of statistical thermodynamics is the Boltzmann distribution. The result states the following: consider a *system* that in *thermal equilibrium* with a *reservoir* at temperature T. Each of the terms in italics have very specific meanings, which will be described shortly. Let E_1 and E_2 denote two energy states of the system. The

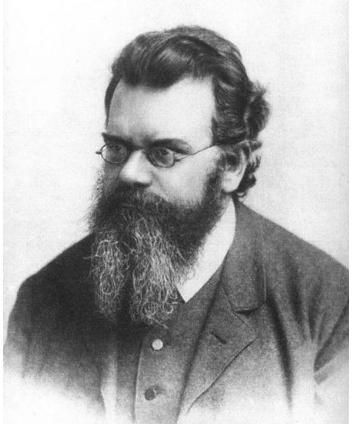


Fig. 4.2: Ludwig Boltzmann, the father of statistical mechanics and kinetic theory of gases. He discovered the formula for entropy or disorder as a mathematical concept: this formula $S = k_B \log\Omega$ is inscribed on his gravestone. The concept of entropy permeates all branches of sciences, including communication systems. Boltzmann ended his life by hanging himself in 1906.

[1] For a detailed derivation, see Thermal Physics by Kittel and Kroemer.

Boltzmann result asserts that the probabilities of finding the system in these energies is related by

$$\boxed{\frac{P(E_1)}{P(E_2)} = \frac{e^{-\beta E_1}}{e^{-\beta E_2}},} \tag{4.1}$$

where $\beta = \frac{1}{k_B T}$, and k_B is the Boltzmann constant. Figure 4.1 illustrates the meanings of the terms in italics. The *reservoir* is a large source of particles and energy, characterized by a temperature T. It goes by the name *reservoir* because it can either take in, or give out any energy without changing its temperature T. As opposed to the reservoir, the *system* is much smaller, and can be found in energy states E_1, E_2, E_3, ... The statement that the system is in *thermal equilibrium* with the reservoir means that it can exchange energy with the reservoir, but *not particles.* Each energy state E_i is considered to be *individually* in thermal equilibrium with the reservoir. Only under this condition is the Boltzmann result in Equation 4.1 applicable. Since the temperature T is the measure of the energy which is being exchanged, the reservoir and the system share the same temperature upon reaching thermal equilibrium.

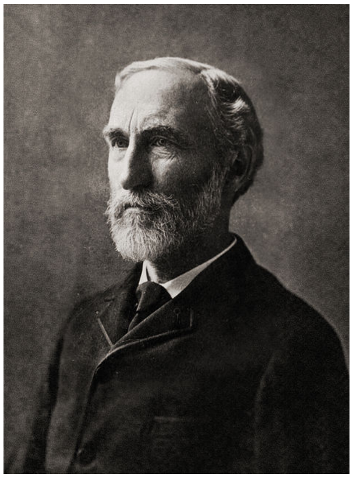


Fig. 4.3: Josiah Willard Gibbs coined the word 'statistical mechanics', and with Maxwell and Boltzmann, gave it its modern rigorous mathematical basis. He developed the modern form of vector calculus. Gibb's quiet lifestyle was in stark contrast to the eternal impact of his scientific ideas and work.

Now if we let the system exchange energy *and particles* with the reservoir, as indicated in Figure 4.1, the Boltzmann relation needs to be generalized. A measure of the particle number is the chemical potential μ, which must also appear in addition to the temperature T in relations characterizing thermodynamic equilibrium between the system and the reservoir. This famous generalization was done by Gibbs, who gave the modified relation

$$\boxed{\frac{P(E_1)}{P(E_2)} = \frac{e^{-\beta(E_1 - n_1\mu)}}{e^{-\beta(E_2 - n_2\mu)}} \underbrace{=}_{\text{non-interacting}} \frac{e^{n_1\beta(\mu - \mathcal{E}_1)}}{e^{n_2\beta(\mu - \mathcal{E}_2)}},} \tag{4.2}$$

where μ is a common chemical potential of the reservoir+system, and n_i is the number of particles in the single-particle energy state $\mathcal{E}_i$. We are going to call a single particle energy eigenstate an *orbital*, drawing from the language of chemistry. Only if the particles considered are *non-interacting*, then the energy of the state is $E_i = n_i\mathcal{E}_i$ if there are n_i particles in orbital $|i\rangle$ of eigenvalue $\mathcal{E}_i$. If these conditions are met, then one defines a Gibbs-sum, or more popularly known as the grand partition function

$$Z = \sum_{\text{states}} \sum_{\text{n}} e^{\beta(n\mu - E_n)}. \tag{4.3}$$

The sum runs over all states of the system, and all number of particles allowed in each single-particle state. Note carefully what this means. For example, consider the situation when orbital $|3\rangle$ is in equilibrium with the reservoir. Since it is not interacting with the other orbitals (which are also separately in equilibrium with the reservoir), the partition function for the 'system' consisting of a variable number of particles in $|3\rangle$ is then $Z = \sum_{n_3=0}^{n_3=n_{max}} e^{\beta n_3(\mu - \mathcal{E}_3)}$. The 'system' here is the various occupation states of orbital $|3\rangle$.

When energy *and* particle exchange is allowed between the system and the reservoir, the fundamental law of equilibrium statistical mechanics may be stated as the following. Under thermodynamic equilibrium with a reservoir at temperature T, the absolute probability that the system will be found in the state $E_i = n_i\mathcal{E}_i$ with n_i particles in orbital $|i\rangle$ is

$$\boxed{P(E_i) = \frac{e^{\beta(n_i\mu - E_i)}}{Z} = \frac{e^{\beta n_i(\mu - \mathcal{E}_i)}}{Z} = \frac{e^{\beta n_i(\mu - \mathcal{E}_i)}}{\sum_{n_i=0}^{n_i=n_{max}} e^{\beta n_i(\mu - \mathcal{E}_i)}}.} \tag{4.4}$$

4.2 Partition Function for Quantum Systems

For sake of completeness and for future use, we generalize the above result for quantum systems. This section may be skipped in an initial reading. We recognize that the allowed orbital energies of any quantum system $\mathcal{E}_i$ are the eigenvalues of the single-particle Hamiltonian $\hat{H}_0$ via $\hat{H}_0|i\rangle = \mathcal{E}_i|i\rangle$, the non-interacting many-particle Hamiltonian $\hat{H} = \sum \hat{H}_0$ gives $\hat{H}|n_1, n_2, ...n_i, ...\rangle = (\sum_i n_i\mathcal{E}_i)|n_1, n_2, ...n_i, ...\rangle$, and the number n_i of particles in the eigenstate (or orbital) $|i\rangle$ is $\hat{N}_i|n_1, n_2, ...n_i, ...\rangle = n_i|n_1, n_2, ...n_i, ...\rangle$, where $\hat{N}_i$ is occupation number operator for eigenstate $|i\rangle$, and $\hat{N} = \sum_i \hat{N}_i$. Then, the expectation value of any operator $\langle\hat{O}\rangle$ at thermodynamic equilibrium is

$$\boxed{\langle\hat{O}\rangle = \frac{\mathrm{Tr}[\hat{O}e^{\beta(\mu\hat{N}-\hat{H})}]}{\mathrm{Tr}[e^{\beta(\mu\hat{N}-\hat{H})}]},} \tag{4.5}$$

where Tr[...] stands for the Trace of the matrix or the operator. Note that the Hamiltonian matrix and the number operator are exponentiated. The Trace gives the sum of the diagonal elements, making Equation 4.5 equivalent to 4.4 in the diagonal representation. But since the Trace is invariant between representations, Equation 4.5 also holds for non-diagonal conditions. Feynman[2] calls the fundamental results in Equation 4.4 (and 4.5) the "summit of statistical mechanics, and the entire subject either a slide-down from the summit, or a climb up to this result". We have not covered the climb-up, but since we will apply the result, let us slide down by applying it to derive the Fermi-Dirac and the Bose-Einstein distribution functions. We will use the version of Equation 4.5 in later chapters, and focus on Equation 4.4 for this chapter.

Fig. 4.4: Richard Feynman is one of the most well-known physicists because of his colorful lifestyle. Made significant contributions to several areas, and developed the path integral approach to quantum mechanics, as a method distinct from Schrodinger and Heisenberg approaches. His contributions to quantum electrodynamics and the eponymous Feynman diagrams won him the 1965 physics Nobel prize.

4.3 The Fermi-Dirac Distribution

As we have discussed in Chapter 3, the number of Fermionic particles that can occupy an energy eigenstate $\mathcal{E}_i$ are $n_i = 0$ or 1 and *nothing else* because of the **Pauli exclusion principle**. Therefore, the partition function for the state of the system corresponding to energy $\mathcal{E}_i$ in thermodynamic equilibrium (in the Gibbs sense) with a reservoir of temperature T and chemical potential μ is simply

$$Z = \sum_{n_i=0}^{n_i=1} e^{\beta n_i(\mu-\mathcal{E}_i)} = e^0 + e^{\beta(\mu-\mathcal{E}_i)} = 1 + e^{\beta(\mu-\mathcal{E}_i)}, \tag{4.6}$$

and the probability that the system is in a state that has n_i particles in orbital $|i\rangle$ is simply $P(E_i) = e^{\beta(n_i\mu - E_i)}/Z$, where $E_i = n_i\mathcal{E}_i$ is the total energy of the orbital. Note that we are assuming that the particles that fill the orbital do not interact with each other. Then, the thermal average number of particles $\langle n_i\rangle$ in orbital $|i\rangle$ is given by $f(\mathcal{E}_i) = \langle n_i\rangle = \sum_i n_i P(E_i)$, which is

$$\langle n_i\rangle = f(\mathcal{E}_i) = \frac{0\cdot e^0 + 1\cdot e^{\beta(1\cdot\mu - 1\cdot\mathcal{E}_i)}}{1+e^{\beta(\mu-\mathcal{E}_i)}} \implies \boxed{f_{FD}(\mathcal{E}_i) = \frac{1}{1+e^{\beta(\mathcal{E}_i-\mu)}}}, \tag{4.7}$$

where the boxed equation is the Fermi-Dirac distribution. Note that it varies between 0 and 1, and is equal to $\frac{1}{2}$ when $\mathcal{E}_i = \mu$. We will discuss this further shortly.

[2] *Statistical Mechanics*, by R. P. Feynman. About his Nobel prize in 1965, Feynman recounts: "I was in the Cornell cafeteria and some guy, fooling around, throws a plate in the air. As the plate went up in the air I saw it wobble, and I noticed the red medallion of Cornell on the plate going around. It was pretty obvious to me that the medallion went around faster than the wobbling. I had nothing to do, so I start figuring out the motion of the rotating plate. I discovered that when the angle is very slight, the medallion rotates twice as fast as the wobble ratetwo to one. It came out of a complicated equation! I went on to work out equations for wobbles. Then I thought about how the electron orbits start to move in relativity. Then there's the Dirac equation in electrodynamics. And then quantum electrodynamics. And before I knew it the whole business that I got the Nobel prize for came from that piddling around with the wobbling plate." A replica of the Cornell plate is now part of an exhibit marking the centennial of the Nobel Prize. Feynman's Messenger lecture series at Cornell are highly receommended, and can be viewed here: http://www.cornell.edu/video/playlist/richard-feynman-messenger-lectures

4.4 The Bose-Einstein Distribution

Unlike Fermions, there is no restriction on the number of Bosonic particles that can occupy an orbital $|i\rangle$. This means $n_i = 0, 1, ..., \infty$. Then, the partition function is

$$Z = \sum_{n_i=0}^{\infty} e^{\beta n_i(\mu - \mathcal{E}_i)} = \sum_{n_i=0}^{\infty} [e^{\beta(\mu-\mathcal{E}_i)}]^{n_i} = \frac{1}{1 - e^{\beta(\mu-\mathcal{E}_i)}}, \tag{4.8}$$

where the infinite sum is a geometric series $1 + u + u^2 + ... = \frac{1}{1-u}$, valid for $u = e^{\beta(\mu-\mathcal{E}_i)} < 1$, or equivalently $\mu \leq \mathcal{E}_i$. The thermal average number of bosonic particles in orbital $|i\rangle$ is then

$$\langle n_i \rangle = f(\mathcal{E}_i) = \frac{0 \cdot u^0 + 1 \cdot u^1 + 2 \cdot u^2 + 3 \cdot u^3 + ...}{(1-u)^{-1}} \implies \boxed{f_{BE}(\mathcal{E}_i) = \frac{1}{e^{\beta(\mathcal{E}_i - \mu)} - 1}}, \tag{4.9}$$

where the boxed equation is the Bose-Einstein distribution. In arriving at the result, we used the relation $u\frac{d}{du}(\frac{1}{1-u}) = \frac{u}{(1-u)^2} = u + 2u^2 + 3u^3 + ...$, which is the sum that appears in the numerator, whereupon $\langle n_i \rangle = \frac{1}{u^{-1}-1}$. Note that for $\beta(\mathcal{E}_i - \mu) >> 1$, the Bose-Einstein distribution $f_{BE}(E_i) \to 0$. However, for $\beta(\mathcal{E}_i - \mu) << 1$, $f_{BE}(E_i) \approx \frac{1}{\beta(\mathcal{E}_i - \mu)}$ can increase without bound, which is surprisingly physical and indicates a *condensation* of all particles to the lowest energy orbitals. This phenomenon is related to Bose-Einstein condensation, a topic to be discussed further later in the book.

4.5 Manifestations of the distribution functions

The key ideas and results in arriving at the distribution functions are summarized in Figure 4.1. In Figure 4.5, we plot the various distribution functions.

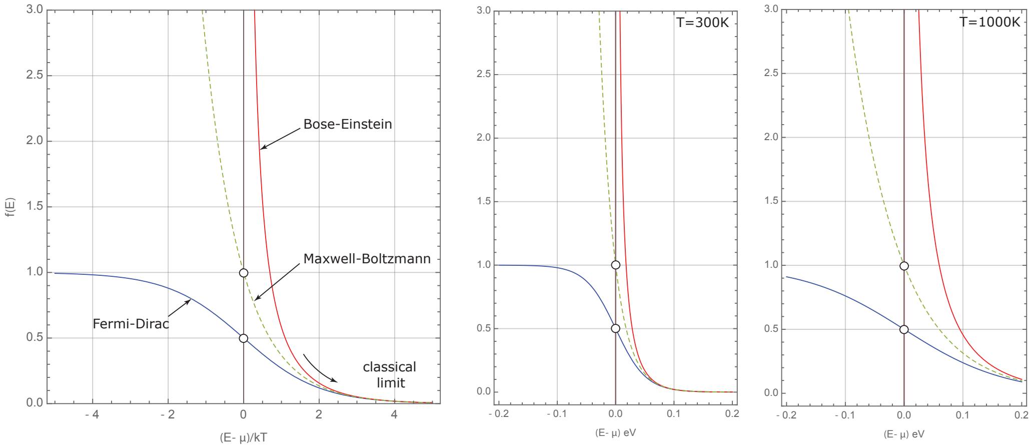


Fig. 4.5: Illustration of the distribution functions and the effect of temperature.

We define the Fermi-Dirac *function* as

$$f_0(x) = \frac{1}{1 + e^{\beta x}} \tag{4.10}$$

which takes the argument $x = E - \mu$ to give us the Fermi-Dirac distribution

$$f_{FD}(E) = f_0(E - \mu) = \frac{1}{1 + e^{\beta(E-\mu)}}. \tag{4.11}$$

The distribution may be thought of a function of the energy E, or of the chemical potential μ. We use the compact notation $f_0 = f_0(E-\mu) = f_{FD}(E)$. The partial derivative with respect to energy is

$$\frac{\partial f_0}{\partial E} = -\frac{\partial f_0}{\partial \mu} = -\beta \cdot \frac{e^{\beta(E-\mu)}}{(1+e^{\beta(E-\mu)})^2} = -\beta \cdot f_0[1-f_0], \tag{4.12}$$

which can be rearranged to the form

$$\boxed{-\frac{\partial f_0}{\partial E} = +\frac{\partial f_0}{\partial \mu} = \frac{\beta}{4\cosh^2(\frac{\beta(E-\mu)}{2})}.} \tag{4.13}$$

The derivative of the Fermi-Dirac distribution evidently reaches its maximum value of $\frac{\beta}{4} = \frac{1}{4kT}$ at $E = \mu$. We have the identity $\int_{-\infty}^{+\infty} du \frac{\beta}{4\cosh^2[\frac{1}{2}\beta u]} = 1$, which indicates that in the limit of very low temperatures $\frac{1}{kT} = \beta \to \infty$, the derivative function should approach a Dirac-delta function in the energy argument, i.e.,

$$\boxed{\lim_{T\to 0}[-\frac{\partial f_0}{\partial E}] = \lim_{T\to 0}[+\frac{\partial f_0}{\partial \mu}] = \delta(E-\mu).} \tag{4.14}$$

This feature is illustrated in Figure 4.6. [3]

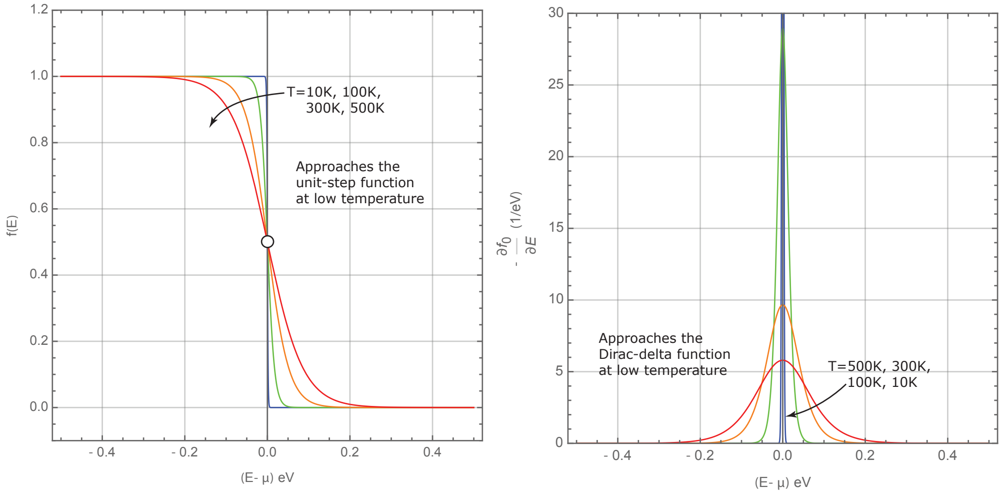


Fig. 4.6: Illustration of the temperature dependence of the Fermi-Dirac distribution, and its derivative.

Now considering $f(u) = 1/(1+e^u)$ and $f(v) = 1/(1+e^v)$, we get the identity

$$\boxed{f(u) - f(v) = \underbrace{[f(u) + f(v) - 2f(u)f(v)]}_{\geq 0} \times \tanh(\frac{v-u}{2})} \tag{4.15}$$

Since $f(u), f(v) \leq 1$, the term in the square brackets is always positive. So the sign of the Fermi difference function is determined by the tanh(...) term. The Fermi difference function will make its appearance repeatedly when we study the optical and electronic transport properties of semiconductors and electronic and photonic devices[4].

The integral of the Fermi-Dirac function is

$$\boxed{\int_0^\infty dE f_0(E-\mu) = \int_0^\infty \frac{dE}{1+e^{\beta(E-\mu)}} = \frac{1}{\beta}\ln(1+e^{\beta\mu}),} \tag{4.16}$$

[3] Because of Pauli Exclusion principle, electrons that contribute to electrical conductivity are in the small window of energies where the Fermi derivative function peaks. Because states of lower energy are completely occupied, and cannot move, and states with energies too high have no electrons. This is the same effect as when we blow air on a bowl of water, the surface responds, not the interior.

[4] For example, the switching of sign of the Fermi difference function will be critical to the creation of population inversion in a LASER.

which leads to the very useful Fermi *difference* integral

$$\boxed{\int_0^\infty dE[f_0(E-\mu_1)-f_0(E-\mu_2)] = \frac{1}{\beta}\ln[\frac{1+e^{\beta\mu_1}}{1+e^{\beta\mu_2}}] = (\mu_1-\mu_2)+\frac{1}{\beta}\ln[\frac{1+e^{-\beta\mu_1}}{1+e^{-\beta\mu_2}}].} \tag{4.17}$$

If $\mu_1, \mu_2 >> kT$, the second term on the rightmost side is zero, and we obtain

$$\boxed{\int_0^\infty dE[f_0(\mu_1)-f_0(\mu_2)] \approx (\mu_1-\mu_2).} \tag{4.18}$$

That this relation is an identity is evident at $T \to 0$, or $\beta \to \infty$. The features of the Fermi difference function are illustrated in Figure 4.7. The integral at low temperatures is just the area under the dashed difference curve, which is rectangular and has a energy width of $\mu_2 - \mu_1$. [5]

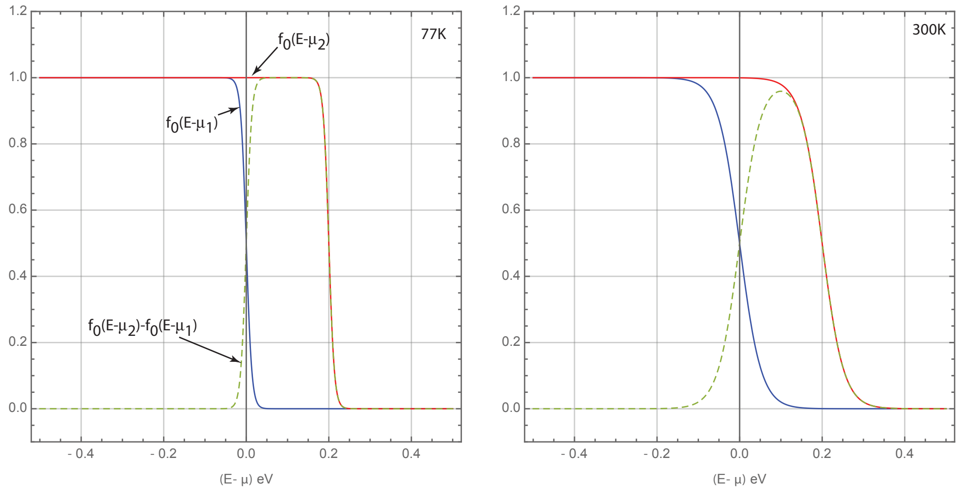


Fig. 4.7: Illustration of the temperature dependence of the Fermi-difference distribution. The difference is a window between $\mu_2 - \mu_1$ that becomes increasingly rectangular as the temperature drops.

It is useful to define higher moment integrals of the Fermi-Dirac functions of the form

$$F_j(\eta) = \frac{1}{\Gamma(j+1)}\int_0^\infty du \frac{u^j}{1+e^{u-\eta}}. \tag{4.19}$$

The Fermi-Dirac integral is rendered dimensionless by scaling the chemical potential $\eta = \beta\mu$, and the energy $u = \beta E$ by the thermal energy $kT = \frac{1}{\beta}$. Since we are integrating over u, the Fermi-Dirac integral $F_j(\eta)$ is a function of the chemical potential μ. The denominator is a normalizing Gamma function $\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx$ with the property $\Gamma(n+1) = n\Gamma(n)$, which means if n is an integer, $\Gamma(n) = (n-1)!$. A useful value of the Gamma function for a non-integer argument is $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. For $\eta << -1$, the exponential in the denominator is much larger than unity. An excellent approximation of the Fermi-Dirac integral then is $F_j(\eta) \approx e^\eta$, irrespective of the value of j. In the other extreme, when $\eta >> 1$, an excellent approximation is $F_j(\eta) \approx \frac{\eta^{j+1}}{\Gamma(j+2)}$. Due to the high importance of Fermi-Dirac integrals in semiconductor devices, we collect the results:

$$\boxed{F_j(\eta) = \frac{1}{\Gamma(j+1)}\int_0^\infty du \frac{u^j}{1+e^{u-\eta}}}, \boxed{F_j(\eta) \underbrace{\approx}_{\eta<<-1} e^\eta}, \boxed{F_j(\eta) \underbrace{\approx}_{\eta>>1} \frac{\eta^{j+1}}{\Gamma(j+2)}}. \tag{4.20}$$

[5] The Fermi difference function will appear later in the problem of electrical current flow. Electrons in a metal or a semiconductor are subjected to two Fermi levels when connected to a battery. The electrons in the Fermi difference function window are those responsible for current flow.

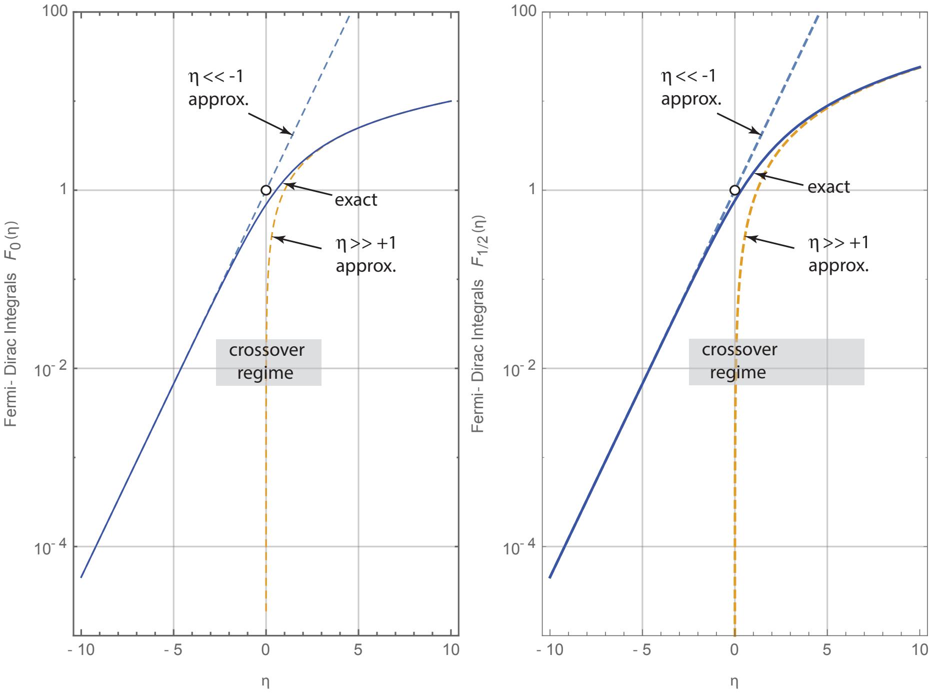


Fig. 4.8: Fermi-Dirac integrals and their non-degenerate ($\eta << -1$) and degenerate ($\eta >> 1$) approximations, illustrating Equation 4.20.

From Equation 4.16, we have an exact analytical result for the Fermi-Dirac integral for $j = 0$: it is $F_0(\eta) = \ln(1 + e^{\eta})$. The validity of the approximations in Equation 4.20 are easily verified for this special case. [6] No exact analytical expressions for other orders ($j \neq 0$) exist. The approximations in Equation 4.20 then assume increased importance for analytical evaluation of various physical quantities such as the mobile carrier densities in semiconductor bands, transport phenomena, and optical properties. The order j depends on the dimensionality of the problem. Figure 4.8 illustrates the cases of the Fermi-Dirac integrals and their approximations for the cases of $j = 0$ and $j = \frac{1}{2}$.

4.6 Meaning of equilibrium in semiconductor devices

You may skip this section in an initial reading. If you are familiar with semiconductors, will find the discussion easygoing and useful. If you are not familiar with semiconductors, please come back to this section after reading about bandgaps and energy-band diagrams. I will refer to this section later when you are ready.

Let us now consider a few semiconductor devices to develop a deeper understanding of the meaning of equilibrium in semiconductor devices. The first and simplest example is a 1D semiconductor (for example a carbon nanotube or a thin nanowire), which has *ohmic contacts* to two metal electrodes. The allowed energy eigenvalues in the semiconductor channel are those in the valence and conduction bands with band edge energies E_v, E_c, separated by a bandgap E_g, as indicated in Figure 4.9. Consider the 1D semiconductor to be doped n-type, with mobile electrons in the conduction band, and no mobile carriers in the valence band. Then the true meaning of an *ohmic contact* is the following: the electrons in the conduction band of the semiconductor are in thermodynamic equilibrium with the electrons in the metal contacts, in the Gibbs-sense. The conduction band states

[6] The Fermi-Dirac integrals in Figure 4.8 are central to understanding the operation of electronic switches, or transistors. The current flowing in a transistor is proportional to a Fermi-Dirac integral in the y-axis, while the voltage controlling the current is proportional to the $x-$axis. As a result, we will see in later chapters that vast, orders of magnitude changes in current can be obtained by small changes in the voltage. The high-current is the on-state, and the low-current will be the off-state of the transistor switch.

(or orbitals) in the semiconductor can freely exchange particles (electrons) and energy with the states or orbitals in the contacts, which is the reservoir. Connect this concept of Gibbs equilibrium in Figure 4.9 with the picture we used earlier in Figure 4.1. Note here we have *two reservoirs*. The particles in the left contact (reservoir) are in equilibrium with each other, and those in the right contact are in equilibrium with each other. When no external voltage is applied across them, the contacts are also in thermodynamic equilibrium with each other.

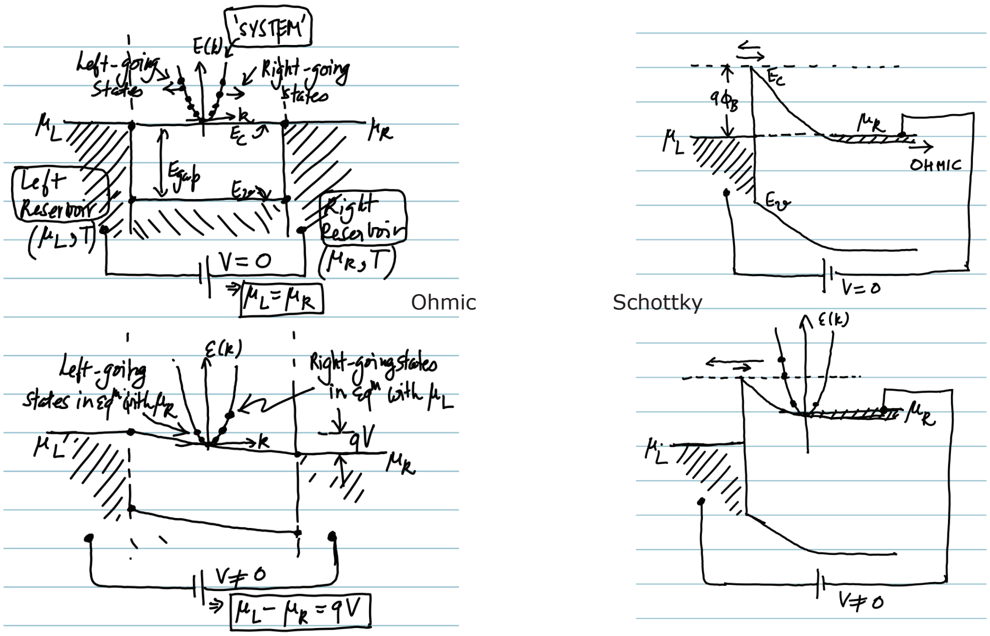


Fig. 4.9: Illustration of the concept of equilibrium for Ohmic and Schottky contacts between metals and semiconductors.

Inside the semiconductor connecting the contacts, there are particles that are moving to the right, and those moving to the left. Let us consider the situation where the left- and right-going carriers *do not mix*, i.e., there is no scattering of carriers. This is referred to as the *ballistic* case, and is approximately realized for very short semiconductor lengths. Consider the electrons moving to the right in the semiconductor. These electrons can only enter the semiconductor from the left contact. Then the electrons moving to the right are in thermodynamic equilibrium with the *left contact*. Similarly, carriers moving to the left in the semiconductor are in equilibrium with the *right contact*. Being in thermodynamic equilibrium in the Gibbs sense means the right-moving electron states share the same chemical potential μ and temperature T as the electrons in the left contact metal. This is an extremely important consequence of thermodynamic equilibrium. Similarly the carriers in the semiconductor moving to the left could only have entered from the right contact, which keeps them in equilibrium with that contact and share μ and T. As long as the chemical potentials of the contacts are the same, the net current flow due the left and right moving carriers in the semiconductor exactly cancel, because they share the same μ.

When a voltage is applied between the contacts, the chemical potential of one contact is $\mu_L - \mu_R = qV$ larger than the other. This in turn breaks the delicate balance of left-and right-moving carriers inside the semiconductor. The imbalance of the left and right moving carriers as indicated in Figure 4.9 thus is the driver of an electric current through the semiconductor, completing the circuit. We will use this picture to calculate the current through a ballistic semiconductor channel in Chapter 10, and show that the conductance is quantized. Then we will use this idea in Chapter 11 in the three-terminal Ballistic transistor switch.

If the chemical potential potential of the metal lines up with energies in the bandgap of the semiconductor, a Schottky contact results, as indicated in Figure 4.9. The figure shows again a semiconductor in contact with two metals: the left contact is now Schottky, and the right contact is ohmic to the conduction band electrons in the semiconductor. Going back to our discussion of equilibrium in Section 4.1, we realize that the left-moving electrons in the semiconductor are in thermodynamic equilibrium with the right contact. But the right moving electrons in the semiconductor are *not* in thermodynamic equilibrium with the left contact in the Gibbs sense, because there is a barrier between them that prevents free particle exchange. When a voltage is applied across the two metal contacts, the stronger imbalance of equilibrium between the left-and right-going carriers in the semiconductor cause a high asymmetry in the current flow as a function of the voltage. For the 'forward' bias condition shown, the left-moving carriers in the semiconductor that make it over the barrier to the metal are in equilibrium with the right contact. Since their chemical potential changes linearly with voltage, their concentration increases exponentially with voltage, causing a characteristic exponential turn-on of the diode. We will discuss this quantitatively in later chapters.

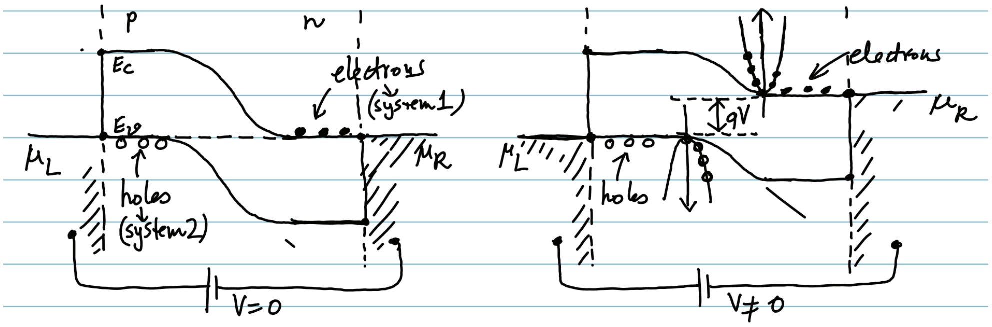


Fig. 4.10: Illustration of the concept of equilibrium for p-n junctions.

Figure 4.10 shows a semiconductor p-n junction. Note the reservoirs are metals, but clearly we have chosen two *different* metals to form ohmic contacts individually to the p-side and the n-side of the semiconductor. An ohmic contact between a semiconductor and one metal electrode is possible for carriers in only one of the semiconductor bands, not both. This means with the proper choice of metals, we can form an ohmic contact to the conduction band of a n-type semiconductor for a n-type ohmic contact, and to the valence band of a p-type semiconductor for a p-type ohmic contact separately. So the holes in the valence band of the p-type semiconductor layer are in thermodynamic equilibrium with the p-ohmic metal (left), and the electrons in the n-type semiconductor layer are in thermodynamic equilibrium with the n-contact metal. Note now we also have two types of carriers - electrons in the conduction band, and holes in the valence band. When no voltage is applied, the holes in the p-side are in thermodynamic equilibrium with the electrons in the n-side - because they are in turn in equilibrium with their respective ohmic contact metal reservoirs. So they share a common chemical potential. However, when a voltage is applied, as indicated in Figure 4.10, the equilibrium is broken; the chemical potentials of the conduction band electrons in the n-side and valence band holes in the p-type now differ by $\mu_n - \mu_p = qV$. This again is responsible for current flow, as will be discussed in later chapters.

As a final example, consider a 3-terminal device, the field-effect transistor (FET) shown in Figure 4.11. The carriers in the inversion channel have ohmic contacts to the source and drain contacts, but a Schottky-type contact through an additional insulating barrier layer to the gate metal reservoir. So the carriers in the semiconductor channgel can be in thermal equilibrium with the carriers in the gate metal, but not in thermodynamic equilibrium in the Gibbs sense because the exchange of particles between the gate metal and the semiconductor channel is explicitly prohibited. The right-going carriers are again injected from the left contact, and the left-going carriers are injected from the right contact. But the carrier density in the semiconductor is controlled by the gate voltage capacitively. We will use this picture in Chapter 11 to discuss the Ballistic FET in detail. The FET is the most commonly used semiconductor device today.

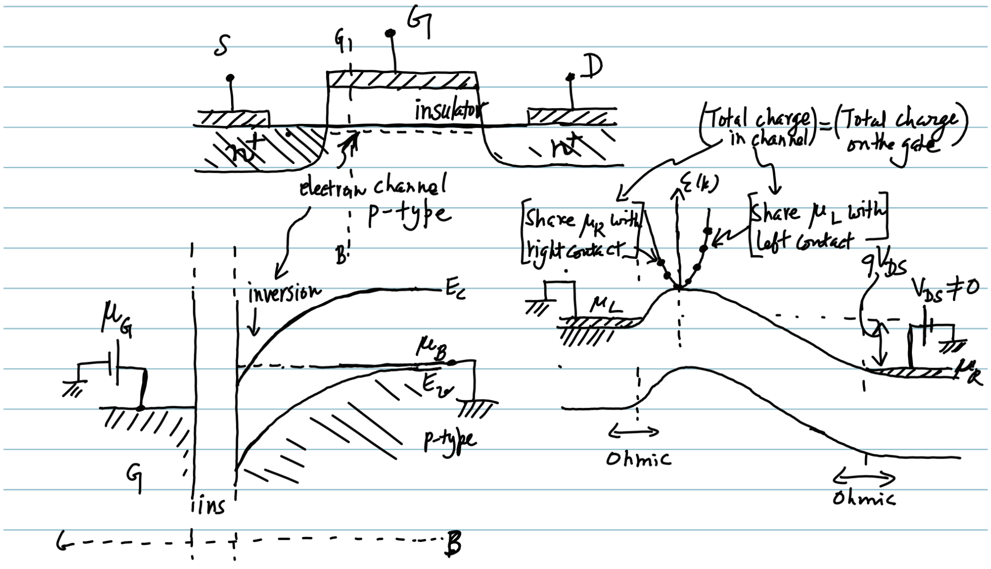


Fig. 4.11: Illustration of the concept of equilibrium for a 3-terminal MOSFET device.

Chapter Summary

- When many indistinguishable **Fermions** such as the electron are allowed to distribute in energy, the occupation function is given by the Fermi-Dirac distribution $\boxed{f_{FD}(\mathcal{E}) = \frac{1}{1 + e^{\beta(\mathcal{E}-\mu)}}}$. Here $\beta = 1/k_B T$ indicates that the electrons are in Gibbs equilibrium with a reservoir of electrons at temperature T, and the number of Fermions (or electrons) determines the chemical potential μ. Pauli exclusion principle clamps the maximum occupation of an orbital to 1.

- When many indistinguishable **Bosons** such as the electron are allowed to distribute in energy, the occupation function is given by the Bose-Einstein distribution $\boxed{f_{BE}(\mathcal{E}) = \frac{1}{e^{\beta(\mathcal{E}-\mu)} - 1}}$. The Bosons of this system are in Gibbs equilibrium with a reservoir of Bosons at temperature T.

- In semiconductors, electrons are the prototype Fermions. The fundamental electronic and photonic properties will be predominantly determined by the Fermi-Dirac function. The Bose-Einstein distribution function will make its appearance when we discuss the interaction of photons with electrons, or of phonons with electrons. Photons and phonons are Bosons. Excitons, which are electron-hole quasiparticles, behave as composite Bosons.

- In semiconductor electronic and photonic devices, the occupation functions are controlled by applying voltages, currents, or shining light.

Problems

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 5

Electrons in the quantum world

Contents

5.1 In Schrodinger's equation we trust 43
5.2 The free electron 44
5.3 Not so free: particle on a ring 45
5.4 The electron steps into a higher dimension: 2D 48
5.5 Electrons in a 3D box . . 49
5.6 Resolving Drude's dilemma in the Quantum World 51
5.7 The particle in a box . . . 51
5.8 The Dirac-Delta Function 51
5.9 The harmonic oscillator . 52
5.10 The Hydrogen atom . . . 52
5.11 Origin of the Elements: The Periodic Table . . 52
5.12 Origin of the Chemical Bond53
5.13 Electrons in a periodic potential: Bloch Theorem 53

In this chapter we subject the electron to the laws of quantum mechanics embodied by the Schrodinger equation, and quantum statistics embodied by the Pauli exclusion principle as indicated by Fig 5.1. We discover that by upgrading classical mechanics and thermodynamics to their quantum versions, we can explain a vast array of experimental facts for which the classical Drude model of the electron failed. In the process, we encounter a few *exactly* solved problems of quantum mechanics. The set of exactly solved problems is precious, because they form a rigorous underpinning on which rests the edifice of condensed matter physics and the physics of semiconductor nanostructures. We introduce the techniques to find the physical dynamics of *single* electrons - such as its momentum, energy, and the current it carries - all now in the significantly updated quantum version where the wave-nature is imposed on the electron from the beginning. Then we find that because of the Pauli exclusion principle, many-electron systems have an 'internal' energy that is both very large and very bewildering, because is has simply no counterpart in classical mechanics. We introduce the concept of the *density of states*, and our bewilderment turns to joy as we realize that we can not only explain the failures of the Drude model, but have discovered a powerful new bag of tricks that explain and *predict* a far richer range of physical behavior of electrons in bulk materials *and* in nanostructures.

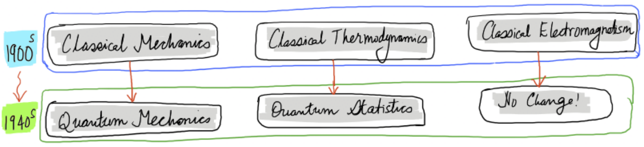


Fig. 5.1: New quantum mechanical rules that govern the behavior of the electron.

5.1 In Schrodinger's equation we trust

As we discussed in Chapter 3, **all physically measurable information** about the quantum states of the electron are buried in the state vector $|\psi\rangle$. By projecting the state vector to the real space we get the wavefunction $\psi(x) = \langle x|\psi\rangle$. In this chapter, we learn how to extract useful information from $\psi(x)$ by applying the corresponding operators of physical observables on them. To do that, we have to first solve the time-independent Schrodinger equation for an electron in various potentials $V(x)$:

$$-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2}\psi(x) + V(x)\psi(x) = E\psi(x). \tag{5.1}$$

The set of solutions $\langle x|n\rangle = \psi(n,x)$ will then be the eigenfunctions corresponding to states of *definite energy* with corresponding eigenvalues E_n. As we learnt in Chapter

3, the **states of definite energy are also stationary states**. They form the most convenient basis for describing the situations when the potential deviates from the ideal, i.e., if $V(x) \to V(x) + W(x,t)$. Thus, the states of definite energy form the basis to uncover what happens when we perturb the quantum system.

In the next few sections we will pick a few potentials $V(x)$ that are central to semiconductor physics. For these potentials, we repeat the following procedure to extract physical information:

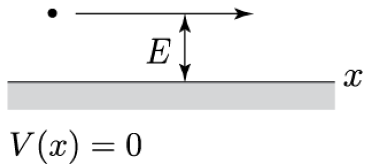
- solve the Schrodinger equation *exactly* to obtain the wavefunction $\psi(x)$ [1],
- the allowed momentum p_x,
- the allowed energy eigenvalues E,
- the quantum current J,
- the Density of States $g(E)$, and
- the *total* energy $\mathcal{U}$, *average* energy u, and energy density u_v of many electrons.

We begin with the simplest of potentials: when $V(x) = 0$.

5.2 The free electron

For $V(x) = 0$, the Schrodinger equation reads

$$-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2}\psi(x) = E\psi(x). \tag{5.2}$$



The equation has the most general solution of the form

$$\boxed{\psi(x) = Ae^{ikx} + Be^{-ikx}}, \tag{5.3}$$

where

$$\boxed{k = \sqrt{\frac{2m_eE}{\hbar^2}} = \frac{2\pi}{\lambda}.} \tag{5.4}$$

We emphasize that the allowed wavelengths λ can take *any* value. Thus, the allowed k values are *continuous*. The allowed energy eigenvalues are

$$\boxed{E(k) = \frac{\hbar^2k^2}{2m_e}}. \tag{5.5}$$

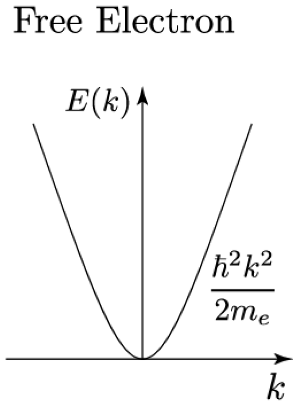


Fig. 5.2: Free electron in 1D. The energy dispersion is parabolic, and *all* k and all $E > 0$ are allowed.

Figure 5.2 shows this parabolic energy dispersion of the free electron. It would not be a stretch to say that this simplest energy dispersion is also one of the most important in all of condensed matter physics. The curvature of the parabola is the inverse mass of the quantum particle. If the particle is heavy, the energies are lower, you can imagine the parabola being pulled down by the heavy mass. Later we will see that the allowed energy eigenvalues of the electron in a semiconductor crystal will develop bands and gaps because of the formation of standing electron waves. Even so, within each electron band, the energy dispersion will again assume a parabolic form at the band edges, but with a different *effective* masses than the free electron mass because of the presence of a periodic crystal potential.

We note that the general solution in Eq. 5.3 represents a superposition of two waves: one going to the right ($\psi_{\rightarrow}(x) = Ae^{ikx}$) and the other to the left ($\psi_{\leftarrow}(x) = Be^{-ikx}$). Since

[1] The Schrodinger equation can be solved exactly only for a very few potentials. We will cover most of them here.

it is a 'mixed' state, clearly it is *not* a state of a definite momentum. We verify this by operating upon the wavefunction by the momentum operator:

$$\hat{p_x}\psi(x) = -i\hbar\frac{d}{dx}\psi(x) = -i\hbar(ikAe^{ikx} - ikBe^{-ikx}) = \hbar k(Ae^{ikx} - Be^{-ikx}) \neq p\psi(x) \quad (5.6)$$

but... for just the right going state we get

$$\hat{p_x}\psi_{\rightarrow}(x) = -i\hbar\frac{d}{dx}\psi_{\rightarrow}(x) = -i\hbar(ikAe^{ikx}) = \hbar k\psi_{\rightarrow}(x) = p\psi_{\rightarrow}(x) \quad (5.7)$$

and it *is* a state of definite momentum. For a rightgoing momentum eigenstate $|+k\rangle$, whose wavefunction is $\psi(x) = Ae^{ikx}$, we find that the quantum charge current density is

$$J(+k) = \frac{q}{2m_e}(\psi^{\star}\hat{p}_x\psi - \psi\hat{p}_x\psi^{\star}) \implies \boxed{J(+k) = q|A|^2\frac{\hbar k}{m_e}}. \quad (5.8)$$

Note that the units are in Amps, because $|A|^2$ has units of 1/length. Similarly, for a left-going state $|-k\rangle$ with wavefunction $\psi(x) = Be^{-ikx}$, the charge current density is

$$J(-k) = \frac{q}{2m_e}(\psi^{\star}\hat{p}_x\psi - \psi\hat{p}_x\psi^{\star}) = -q|B|^2\frac{\hbar k}{m_e}. \quad (5.9)$$

From an analogy to the 'classical' charge current density $J = qnv$, where $n \sim |A|^2$ or $n \sim |B|^2$ is the particle density[2], we identify that the state $|+k\rangle$ has a velocity $\frac{\hbar k}{m_e}$, and the mirror-reflected state $|-k\rangle$ has a velocity $-\frac{\hbar k}{m_e}$. Just like in classical mechanics, the velocity seems to be proportional to the *slope* of the energy dispersion curve $E(k)$. In classical mechanics of particles, the kinetic energy is $E = p^2/(2m)$, and the velocity is $v = dE/dp$. In quantum mechanics, the particle has a wave-like nature, and by analogy we cautiously define the **group velocity** of a quantum particle as

$$\boxed{\mathbf{v}_g(\mathbf{k}) = \nabla_{\mathbf{p}}E(\mathbf{p}) = \frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k})}. \quad (5.10)$$

We suspend further discussion of this definition till later. Using the group velocity, we can write the charge current as $J(k) = qv_g(k)f(k)$, where $f(k)$ is the occupation probability of state $|k\rangle$. Using this procedure, we can find the quantum charge current carried by any superposition state $|\psi\rangle = \sum_k A_k|k\rangle$ we can cook up.

The free electron wavefunction *cannot* be normalized, because it extends over all space from $-\infty \leq x \leq +\infty$. To normalize it, we wrap the infinitely long line and join the infinities to form a circle. Physical quantities that have to do with density, such as the 'density of states' or the 'energy density' of the completely free electron are ill defined because of the infinite volume it lives in. So we first put the electron in a circular ring to calculate these quantities.

5.3 Not so free: particle on a ring

Figure 5.3 shows an electron restricted to move on a circular ring of circumference L, with $V(x) = 0$. Though it is not exactly a 1D problem, we assign one linear coordinate x to the particle's location. We demand all solutions to the Schrodinger equation to be *single-valued* functions of x. Because the loop closes on itself, the electron wavefunction must satisfy[3]

$$\psi(x+L) = \psi(x) \rightarrow e^{ik(x+L)} = e^{ikx} \rightarrow e^{ikL} = 1 \rightarrow kL = 2n\pi \quad (5.11)$$

to be single-valued. This is only possible if

[2] If you are uncomfortable with this statement, I am with you. $n \sim |A|^2$ is true *only if* the particle is confined, as we will see in the next section. The completely free electron wavefunction is not normalizable!

[3] This periodic boundary condition also is referred to as the Born von-Karman boundary condition. It is mathematically distinct from the 'hard-wall' boundary condition what we will impose for the particle in a box problem, but the physics of the interior will not be affected by this choice.

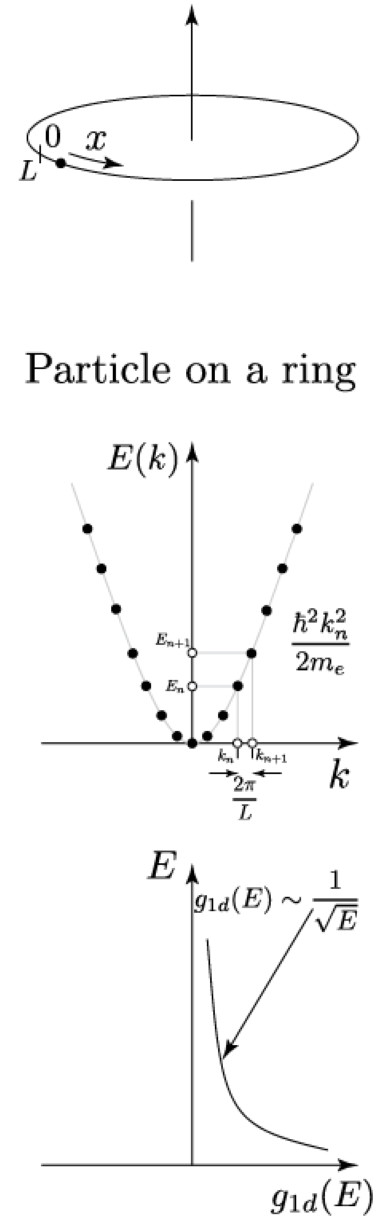


Fig. 5.3: Putting the electron on a ring *quantizes* the allowed wavevectors k_n, and as a result the momentum, the angular momentum, and the energy of the particle are quantized. The density of states for the electron on a ring with parabolic energy dispersion goes as $1/\sqrt{E}$, counting how the allowed eigenvalues distribute in energy if we were to put many electrons on the ring.

$$\boxed{k_n = \frac{2\pi}{L} n}, n = 0, \pm 1, \pm 2, ... \tag{5.12}$$

where $\psi_n(x) = Ae^{ik_n x}$. We see that for the particle on a ring, the set of allowed k_n are *discrete* as indicated in Figure 5.2, and thus the allowed momentum are discrete:

$$\boxed{p_n = \hbar k_n = \frac{h}{2\pi}\frac{2\pi}{L} n = n\frac{h}{L}}, \tag{5.13}$$

and the allowed values of the momentum are *quantized.* The smallest spacing of the allowed wavevectors is precisely $\Delta k = k_{n+1} - k_n = 2\pi/L$.

Because the angular momentum is $\mathbf{L} = \mathbf{r} \times \mathbf{p}$, we find that

$$\mathbf{L}_n = \mathbf{r} \times \mathbf{p} = \hbar k_n \times \frac{L}{2\pi}\hat{\mathbf{z}} = \frac{2\pi\hbar}{L} n \times \frac{L}{2\pi}\hat{\mathbf{z}} \implies \boxed{L_n = n\hbar}, \tag{5.14}$$

i.e. like the linear momentum, the *angular* momentum of the electron on a ring is also quantized, and can only take values $..., -2\hbar, -\hbar, 0, +\hbar, +2\hbar, ...$ We gain a physical intuition of what Planck's constant $\hbar$ actually means - it is a measure of the angular momentum of a quantum particle. For example, if I tie a 1 kg mass to a 1 m string and spin it at 1 m/s, the angular momentum is $L_{cl} = 1$ J·s. So for this classical situation, I will be providing the mass $n \sim 10^{34}$ quanta - and I may feel like Superman in the quantum world. But what this example really tells us is precisely *how small a quantum of angular momentum actually is*!

As promised, the eigenfunctions of the particle on a ring can be normalized:

$$\int_0^L dx|\psi_n(x)|^2 = 1 \to |A|^2 \times L = 1 \to A = \frac{1}{\sqrt{L}} \to \boxed{\psi_n(x) = \frac{1}{\sqrt{L}} e^{ik_n x}} \tag{5.15}$$

Note that $n = 0$ is *allowed* as a result of the periodic boundary condition. We observe that the set of functions $[..., \psi_{n-1}(x), \psi_n(x), \psi_{n+1}(x), ...]$ are mutually orthogonal because $\langle m|n\rangle = \int_0^L dx\psi_m^\star(x)\psi_n(x) = \int_0^L dx \frac{e^{i\frac{2\pi}{L}(n-m)x}}{L} = 0$ for $n \neq m$, and $= 1$ for $n = m$. Because the states are orthogonal, we can write

$$\langle m|n\rangle = \int_0^L dx \langle m|x\rangle\langle x|n\rangle = \int_0^L dx\psi_m^\star(x)\psi_n(x) = \delta_{n,m}. \tag{5.16}$$

We also make a note that this set of linearly independent functions is *complete*, meaning if you give me *any* function $f(x)$ in 1D, I can write the function as $f(x) = \sum c_n\psi_n(x)$ with suitable coefficients c_n. This is what we mean by linear superposition: the amazing thing about quantum mechanics is the electron on the ring is allowed to be in *any* state that is a linear superposition of the eigenstates.

The allowed energy eigenvalues are

$$\boxed{E_n = \frac{\hbar^2 k_n^2}{2m_e} = n^2 \frac{(2\pi\hbar)^2}{2m_e L^2} = n^2 \frac{h^2}{2m_e L^2}}. \tag{5.17}$$

The energy eigenvalues are also *quantized*, and grow as n^2. Because the electron is allowed to be in the $n = 0$ state, the minimum energy allowed is $E = 0$. This will *not* be the case if we put the particle in a box in section 5.7. Two important physical intuition we should take away from this example are:

- The smaller the circle, the larger the allowed energies ($L \downarrow \implies E_n \uparrow$), and
- The smaller the mass, the larger the allowed energies ($m \downarrow \implies E_n \uparrow$).

The first statement is one of *quantum confinement*: the smaller the space we fit a quantum particle into, the larger will be its energy. This is because the wavelength must become very small to fit in the space, which means high $k = 2\pi/\lambda$, and higher energy.

A glance at the density of energy eigenvalues in Figure 5.3 shows that they are more densely spaced at low energies, and become sparse at higher energies. We can guess that the dependence with energy must go as $1/E^\eta$, where $\eta > 0$. To find the 1D density of states quantitatively, we note that between $k \to k + dk$, there are $\frac{dk}{\frac{2\pi}{L}}$ allowed states. The total state density $G_{1d}(E)$ in energy is then

$$g_s g_v \frac{2dk}{\frac{2\pi}{L}} = G_{1d}(E)dE \implies g_{1d}(E) = \frac{G_{1d}}{L} = \frac{2g_s g_v}{2\pi\frac{dE}{dk}} \implies \boxed{g_{1d}(E) = \frac{2g_s g_v}{2\pi}(\frac{2m_e}{\hbar^2})^{\frac{1}{2}}\frac{1}{\sqrt{E}}}. \tag{5.18}$$

Because in 1D, the electron could be moving clockwise or counterclockwise, we use $2dk$ to account for the two dk's for $+k$ and $-k$ wavevectors. We have introduced a *spin degeneracy* g_s, which is typically $= 2$ for up and down spins, and a *valley degeneracy* which is the number of copies of such parabolic dispersions that may be present in the $k-$space. Till we introduce crystals, $g_v = 1$, and we assume $g_s = 2$ till we need to consider situations where it does not hold. We note that the 1D DOS decreases as $1/\sqrt{E}$, and has a singularity as $E \to 0$.

Now if instead of a single electron, we fill the ring with N electrons, what would be the total energy of the ensemble? If there are N electrons in the ring, their 1D density is $n = N/L$. We will first completely neglect the Coulomb interaction between the electrons. Though this sounds like heresy, let me assure you that it is actually OK[4]. Besides, the discussion here is equally applicable to neutrons - which do not have charge. Because electrons are Fermions, the moment we go to two, Pauli exclusion principle kicks in; the electron occupation function must follow the Fermi-Dirac distribution $f(E)$. Let us look at the $T = 0$ K situation, when $f(E) = 1$ for $0 \leq E \leq E_F$ and 0 elsewhere. The electrons then must fill up to a Fermi wavevector k_F and Fermi level E_F such that

$$\frac{g_s g_v \times 2k_F}{\frac{2\pi}{L}} = N \implies \boxed{k_F = \frac{\pi}{2}n} \implies \boxed{E_F = \frac{\hbar^2\pi^2 n^2}{8m_e}}. \tag{5.19}$$

This is a remarkable result: the *ground state* of the electron ensemble, at $T = 0$ K, already has a large amount of energy. For example, if we have $n \sim 10^8$/cm which is typical for a metal, then the electron states with the highest energy have $\lambda_F \sim 0.4$ nm and $E_F \sim 10$ eV. This is the energy picked up by an electron in a 10 Volt potential [5]. If we were to provide this energy in the form of heat, $k_B T = E_F$ would lead to $T \sim 10^5$ K. Where did all this energy come from?

This root of this energy reserve is the Fermionic nature of the electron, and the Pauli exclusion principle. There is simply no classical explanation of this energy, it is of a pure quantum origin. We will shortly see that the very high conductivity of metals even at the lowest temperatures is a direct result of these high-energy electrons. And the electrons at the highest energy are *fast*: the Fermi velocity is $v_F = \frac{\hbar k_F}{m_e} = \frac{hn}{4m_e} \sim 5 \times 10^7$ cm/s.

The total energy $\mathcal{U}$ of the ensemble of electrons of density n at $T = 0$ K, and the average energy $u = \mathcal{U}/N$ are then given by

$$\mathcal{U} = \int_0^{E_F} dE \cdot E \cdot G_{1d}(E) \implies \boxed{u_{1d} = \frac{\mathcal{U}}{N} = \frac{\int_0^{E_F} dE \cdot E \cdot G_{1d}(E)}{\int_0^{E_F} dE \cdot G_{1d}(E)} = \frac{1}{3}E_F}, \tag{5.20}$$

and the energy density per length $u_v = \frac{\mathcal{U}}{L}$ is then given by $\boxed{u_v(1d) = \frac{1}{3}nE_F}$. That the

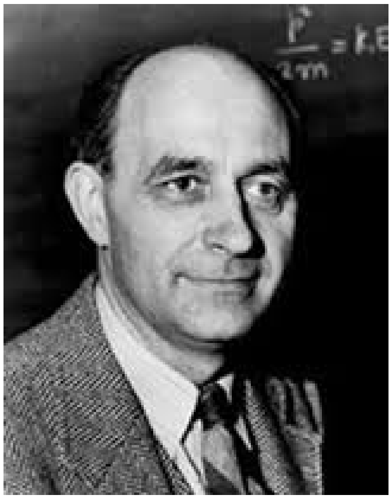


Fig. 5.4: Enrico Fermi, a towering figure whose contributions extended into all areas of physics. Nobel prize recipient in 1938.

[4] The justification came from Lev Landau, and goes under the umbrella of what is called the Fermi-liquid theory. We will encounter it at a later point.

[5] If we pack even more Fermions in small volumes, the Fermi energy is much larger. For example, because of the tight packing of Fermions - protons and neutrons - in the nucleus of an atom, the Fermi energy reaches 40 MeV - yes, *Mega* eV!

average energy of the 1D electron distribution $\frac{1}{3}E_F$ is less than $\frac{1}{2}E_F$ is easily understood from the shape of the DOS $g_{1d}(E)$, which is weighted heavier for lower energies.

5.4 The electron steps into a higher dimension: 2D

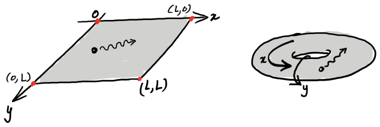


Fig. 5.5: Periodic boundary conditions in 2D leads to a Torus.

Now let the electron move in two dimensions - say in a square box of side L and area $A = L^2$. The spatial coordinate is now a vector $\mathbf{r} = (x, y)$, and the wavevector $\mathbf{k} = (k_x, k_y)$. Learning from the 1D particle on a ring, by subjecting the electron to the Schrodinger equation in two dimensions with $V(\mathbf{r}) = 0$, we find that the allowed wavefunctions are

$$\boxed{\psi(\mathbf{r}) = \frac{1}{\sqrt{L^2}} e^{i(k_x x + k_y y)} = \frac{1}{\sqrt{A}} e^{i\mathbf{k}\cdot\mathbf{r}}}. \tag{5.21}$$

The periodic boundary condition in the **k**-space leads to a torus (Figure 5.5). This leads to the allowed wavevectors

$$\boxed{\mathbf{k} = (k_{n_x}, k_{n_y}) = \frac{2\pi}{L}(n_x, n_y)} \implies \boxed{\mathbf{p} = \hbar\mathbf{k}, |\mathbf{p}| = \frac{h}{L}\sqrt{n_x^2 + n_y^2}}, \tag{5.22}$$

where n_x, n_y are *independent* integers $..., -2, -1, 0, 1, 2, ...$ In the $\mathbf{k}-$space, the allowed set of points form a rectangular grid, each of area $(\frac{2\pi}{L})^2$. Each point in this grid defines an allowed state for electrons. The allowed energy of each such point is

$$\boxed{E(k_x, k_y) = \frac{\hbar^2}{2m_e}(k_{n_x}^2 + k_{n_y}^2) = E(n_x, n_y) = (n_x^2 + n_y^2)\frac{h^2}{2m_e L^2} = \frac{\hbar^2|\mathbf{k}|^2}{2m_e}}. \tag{5.23}$$

To find the 2D DOS $g_{2d}(E)$, we try our intuition like in the 1D case, but we must be careful. Indeed the energies bunch up as we approach $(k_x, k_y) = (0, 0)$, but we must not forget that unlike the 1D case where there were mere two points, we have an entire *circle* of equal-energy states in 2D. In the 2D case, these effects cancel out, giving us a DOS that is *independent* of electron energy:

$$g_s g_v \frac{2\pi k dk}{(\frac{2\pi}{L})^2} = G_{2d}(E)dE \implies \frac{G_{2d}(E)}{L^2} = \boxed{g_{2d}(E) = \frac{g_s g_v m_e}{2\pi\hbar^2}\Theta(E)}. \tag{5.24}$$

Here $2\pi k dk$ is the area of the thin ring of thickness dk around the circle of radius k. Because each state occupies an area $(\frac{2\pi}{L})^2$ in the $\mathbf{k}-$space, there are $\frac{2\pi k dk}{(\frac{2\pi}{L})^2}$ energy states in the ring, from which we get the DOS $g_{2d}(E)$. We also note that because for a free electron, $g_s = 2$ and $g_v = 1$, the 2D DOS is typically written as $g_{2d}(E) = \frac{m_e}{\pi\hbar^2}$ for $E > 0$.

The fact that the 2D DOS for a parabolic energy dispersion is a constant in energy plays a very important role in semiconductor field-effect transistors, where the conducting channel is a 2D electron gas. Moving to the many-electron picture in 2D, let us put N non-interacting electrons in the area A so that the density per unit area is $n = N/A$. At $T = 0$ K, we apply the Pauli exclusion principle again and find that we must fill the states from the center $\mathbf{k} = (0, 0)$ to a sharply defined **Fermi circle** of radius k_F given by

$$g_s g_v \frac{\pi k_F^2}{(\frac{2\pi}{L})^2} = N \implies \boxed{k_F = \sqrt{\frac{4\pi n}{g_s g_v}}}. \tag{5.25}$$

If $g_s = 2$ and $g_v = 1$, we the expression for the Fermi wavevector is $k_F = \sqrt{2\pi n}$. For example, in semiconductor field-effect transistors, typical sheet densities of 2D electron gases (2DEGs) is $n \sim 10^{12}/\text{cm}^2$. The Fermi wavevector is then $k_F \sim 2.5 \times 10^8/\text{m}$, implying a Fermi wavelength of $\lambda_F = \frac{2\pi}{k_F} \sim 25$ nm. On the other hand, for a 2D metal with $n \sim 10^{16}/\text{cm}^2$, the Fermi wavelength is much shorter, $\lambda_F \sim 0.25$ nm.

For non-zero temperatures, the smearing of the Fermi-Dirac distribution near $E = E_F$ makes the Fermi-circle diffuse. To get the electron density in a 2DEG at any temperature T, we use the 2D DOS and use the fact that the electron density does not change with temperature to get the Fermi level E_F thus:

$$n = \int_0^\infty dE \cdot g_{2d}(E) \cdot f(E) = \frac{g_s g_v m_e k_B T}{2\pi\hbar^2} \ln(1 + e^{\frac{E_F}{k_B T}}) \implies \boxed{E_F = k_B T \ln(e^{\frac{n}{n_q}} - 1)}, \tag{5.26}$$

where we have defined a 'quantum' 2D electron concentration $n_q = \frac{g_s g_v m_e}{2\pi\hbar^2} k_B T$. If the temperature is low enough so that $n >> n_q$, then the Fermi level is given simply by $E_F \sim \frac{n}{g_{2d}}$, or $n = g_{2d} E_F$. The meaning of this relation: electron density = (DOS) × (Energy window).

We can then obtain the total energy $\mathcal{U}$ and the average electron energy u at $T = 0$ K as

$$\mathcal{U} = \int_0^\infty dE \cdot E \cdot G_{2d}(E) \cdot f(E) \implies \boxed{u_{2d} = \frac{\mathcal{U}}{N} = \frac{\int_0^\infty dE \cdot E \cdot G_{2d}(E) \cdot f(E)}{\int_0^\infty dE \cdot G_{2d}(E) \cdot f(E)} = \frac{1}{2} E_F}, \tag{5.27}$$

which could have been guessed without the math: because of the constant DOS, the average energy is exactly half the maximum energy E_F at $T = 0$ K. The temperature-dependent average energy requires a bit more work, which we will do for the 3D electron gas. The energy density per unit area at $T = 0$ K is then given by $\boxed{u_v(2d) = \frac{1}{2} n E_F}$.

Before we move to the 3D electron gas, we discuss the quantum current carried by the 2DEG. The wavefunction of a state (n_x, n_y) is $\psi(\mathbf{k}, \mathbf{r}) = \frac{1}{\sqrt{A}} e^{i\mathbf{k}\cdot\mathbf{r}}$. The quantum current due to this state can be obtained from the wavefunction using Equation 3.17:

$$\mathbf{J}(\mathbf{k}) = \frac{q}{2m_e}(\psi^\star \hat{\mathbf{p}} \psi - \psi \hat{\mathbf{p}} \psi^\star) = q \cdot \frac{1}{A} \cdot \frac{\hbar\mathbf{k}}{m_e} = q(\frac{1}{A}) \mathbf{v}_g(\mathbf{k}). \tag{5.28}$$

We can recognize the group velocity appear in the expression of the current. Note that the unit of 2D current density is Amp/m, or current per unit width. A state $|\mathbf{k}\rangle$ has a group velocity $\mathbf{v}_g(\mathbf{k}) = \frac{\hbar\mathbf{k}}{m_e}$ pointing radially outwards from the origin in the $\mathbf{k}-$space. At thermal equilibrium, the occupation of $\mathbf{k}-$space is symmetric around the origin. Thus, for every carrier moving radially outward in one direction, there is an exactly equal current in the opposite direction, which means the net current is zero. But the individual currents carried by the $\mathbf{k}$-states are substantial!

To find the total current of all occupied electron states, we can sum the currents carried by each $\mathbf{k}-$state:

$$\boxed{\mathbf{J} = g_s g_v \sum_{\mathbf{k}} q \mathbf{v}_g(\mathbf{k}) f(\mathbf{k}) = g_s g_v q \int \frac{d^2k}{(2\pi)^2} \mathbf{v}_g(\mathbf{k}) f(\mathbf{k})}. \tag{5.29}$$

5.5 Electrons in a 3D box

The electron is now in a cubic box of side L and volume $V = L^3$ with a spatial coordinate vector $\mathbf{r} = (x, y, z)$, and a wavevector $\mathbf{k} = (k_x, k_y, k_z)$. For $V(\mathbf{r}) = 0$, the allowed wavefunctions are

$$\boxed{\psi(\mathbf{r}) = \frac{1}{\sqrt{L^3}} e^{i(k_x x + k_y y + k_z z)} = \frac{1}{\sqrt{V}} e^{i\mathbf{k}\cdot\mathbf{r}}}. \tag{5.30}$$

The allowed wavevectors are

$$\boxed{\mathbf{k} = (k_{n_x}, k_{n_y}, k_{n_z}) = \frac{2\pi}{L}(n_x, n_y, n_z)} \implies \boxed{\mathbf{p} = \hbar\mathbf{k}, |\mathbf{p}| = \frac{h}{L}\sqrt{n_x^2 + n_y^2 + n_z^2}}, \quad (5.31)$$

where n_x, n_y, n_z are again *independent* integers $..., -2, -1, 0, 1, 2, ...$ In the $\mathbf{k}-$space, the allowed set of points now form a 3D grid, each of volume $(\frac{2\pi}{L})^3$. The allowed energy of each such allowed point is

$$\boxed{E(k_x, k_y, k_z) = \frac{\hbar^2}{2m_e}(k_{n_x}^2 + k_{n_y}^2 + k_{n_z}^2) = (n_x^2 + n_y^2 + n_z^2)\frac{h^2}{2m_e L^2} = \frac{\hbar^2|\mathbf{k}|^2}{2m_e}}. \quad (5.32)$$

The 3D DOS $g_{3d}(E)$:

$$g_s g_v \frac{4\pi k^2 dk}{(\frac{2\pi}{L})^3} = G_{3d}(E)dE \implies \frac{G_{3d}(E)}{L^3} = \boxed{g_{3d}(E) = \frac{g_s g_v}{4\pi^2}(\frac{2m_e}{\hbar^2})^{\frac{3}{2}}\sqrt{E}}. \quad (5.33)$$

Here $4\pi k^2 dk$ is the volume of the thin shell of thickness dk around the sphere of radius k. Because each state occupies a volume $(\frac{2\pi}{L})^3$ in the $\mathbf{k}-$space, there are $\frac{4\pi k^2 dk}{(\frac{2\pi}{L})^3}$ energy states in the shell.

Because the 3D volume increases as k^3, there are more energy states at higher energies; the $g_{3d}(E) \sim \sqrt{E}$ increase in the 3D DOS is a characteristic feature of 3D electron systems with parabolic dispersion: a result well worth remembering. For the many-electron picture in 3D, we put N non-interacting electrons in the volume V so that the density per unit area is $n = N/V$. At $T = 0$ K, the Pauli exclusion principle suggests states must fill from the center $\mathbf{k} = (0, 0, 0)$ to a sharply defined **Fermi Sphere** of radius k_F given by

$$g_s g_v \frac{\frac{4}{3}\pi k_F^3}{(\frac{2\pi}{L})^3} = N \implies \boxed{k_F = (\frac{6\pi^2 n}{g_s g_v})^{\frac{1}{3}}}. \quad (5.34)$$

If $g_s = 2$ and $g_v = 1$, the Fermi wavevector is $k_F = (3\pi^2 n)^{\frac{1}{3}}$. In a metal with $n \sim 10^{24}/\text{cm}^3$, the Fermi wavevector is $k_F \sim 3 \times 10^{10}$/m, and the Fermi wavelength is $\lambda_F \sim 0.2$ nm. The Fermi surface of the 3D electron gas is the surface of this Fermi sphere. It holds the secrets to most of its properties. For free electrons in 3D, the Fermi surface is spherical. When we introduce atoms in a crystal, the surface will deform and assume a rich range of shapes.

For non-zero temperatures, the smearing of the Fermi-Dirac distribution near $E = E_F$ makes the Fermi-circle diffuse. The electron density at a temperature T is constant:

$$n = \int_0^\infty dE \cdot g_{3d}(E) \cdot f(E) = \frac{g_s g_v}{4\pi^2}(\frac{2m_e}{\hbar^2})^{\frac{3}{2}} \int_0^\infty dE \cdot \sqrt{E} \cdot f(E) = n_{3d} F_{\frac{1}{2}}(\eta), \quad (5.35)$$

where the dimensionless Fermi-Dirac integral $F_j(\eta)$ is used to define an *effective* 3D DOS $n_{3d} = \frac{g_s g_v}{4\pi^2}(\frac{2m_e k_B T}{\hbar^2})^{\frac{3}{2}}$, and $\eta = E_F/k_B T$.

The total energy $\mathcal{U}$ and the average electron energy u at $T = 0$ K is

$$\mathcal{U} = \int_0^\infty dE \cdot E \cdot G_{3d}(E) \cdot f(E) \implies \boxed{u_{3d} = \frac{\mathcal{U}}{N} = \frac{\int_0^\infty dE \cdot E \cdot G_{3d}(E) \cdot f(E)}{\int_0^\infty dE \cdot G_{3d}(E) \cdot f(E)} = \frac{3}{5}E_F}. \quad (5.36)$$

Because the DOS increases with E, the average energy is above $\frac{1}{2}E_F$ at $T = 0$ K. The energy density per unit area at $T = 0$ K is then given by $\boxed{u_v(3d) = \frac{3}{5}nE_F}$.

The temperature-dependent average energy requires a bit more work. This was first done famously by Arnold Sommerfeld, who obtained the result for the 3D Fermi gas $\boxed{E_F(T) \approx E_F[1 - \frac{1}{3}(\frac{\pi}{2}\frac{k_bT}{E_F})^2]}$ for $k_BT << E_F$.

The quantum current carried by the 3DEG is To find the total current of all occupied electron states, we can sum the currents carried by each $\mathbf{k}-$state:

$$\boxed{\mathbf{J} = g_s g_v \sum_{\mathbf{k}} q\mathbf{v}_g(\mathbf{k})f(\mathbf{k}) = g_s g_v q \int \frac{d^3k}{(2\pi)^3}\mathbf{v}_g(\mathbf{k})f(\mathbf{k})}. \tag{5.37}$$

which has units of Amp/m^2, or current per unit area. Similar to the 2D case, a state $|\mathbf{k}\rangle$ has a group velocity $\mathbf{v}_g(\mathbf{k}) = \frac{\hbar\mathbf{k}}{m_e}$ pointing radially outwards from the origin in the $\mathbf{k}-$space, and net current at thermal equilibrium is zero because for every $|+\mathbf{k}\rangle$ state, there is a corresponding $|-\mathbf{k}\rangle$ state. This delicate balance is broken when an electric field is applied, which we turn to next.

Fig. 5.6: Arnold Sommerfeld first introduced elliptical orbits to Bohr's model of the atom. Introduced the quantum electron theory of for metals, and resolved the discrepancies of the Drude model. Was the advisor and mentor of a large cohort of Nobel prize winners, but was never awarded the prize in spite of being nominated ~80 times!

Following text till the left arrows is work in progress → →.

5.6 Resolving Drude's dilemma in the Quantum World

5.7 The particle in a box

$$V(x) = 0, \quad 0 \leq x \leq L \tag{5.38}$$

$$V(x) = \infty, \quad x < 0, x > L \tag{5.39}$$

The major change is that $\psi(x) = 0$ in regions where $V(x) = \infty$.

$$\psi(x) = Ae^{ikx} + Be^{-ikx} \rightarrow \psi(0) = 0 = A + B, \psi(L) = Ae^{ikL} + Be^{-ikL} = 0 \tag{5.40}$$

$$\frac{A}{B} = -e^{-i2kL} = -1 \rightarrow 2kL = 2n\pi \rightarrow \boxed{k_n = n\frac{\pi}{L}}, n = \pm1, \pm2, \pm3, ... \tag{5.41}$$

Note that $n = 0$ is *not allowed*, because then $\psi(x) = 0$ and there is no particle. The wavefunction after normalization over the length L is

$$\psi(n, x) = \sqrt{\frac{2}{L}}\sin(n\frac{\pi}{L}x) = \sqrt{\frac{2}{L}}\sin(k_n x) \tag{5.42}$$

$$E_n = n^2\frac{(\pi\hbar)^2}{2m_eL^2} = n^2\frac{h^2}{8m_eL^2} \tag{5.43}$$

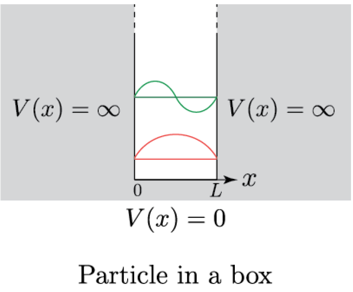


Fig. 5.7: Electron in a box.

5.8 The Dirac-Delta Function

To be written.

5.9 The harmonic oscillator

$$V(x) = \frac{1}{2}m_e\omega^2x^2 \tag{5.44}$$

$$E_n = (n+\frac{1}{2})\hbar\omega \tag{5.45}$$

$$a = \sqrt{\frac{m\omega}{2\hbar}}(\hat{x} + \frac{i}{m\omega}\hat{p}) \tag{5.46}$$

$$a^\dagger = \sqrt{\frac{m\omega}{2\hbar}}(\hat{x} - \frac{i}{m\omega}\hat{p}) \tag{5.47}$$

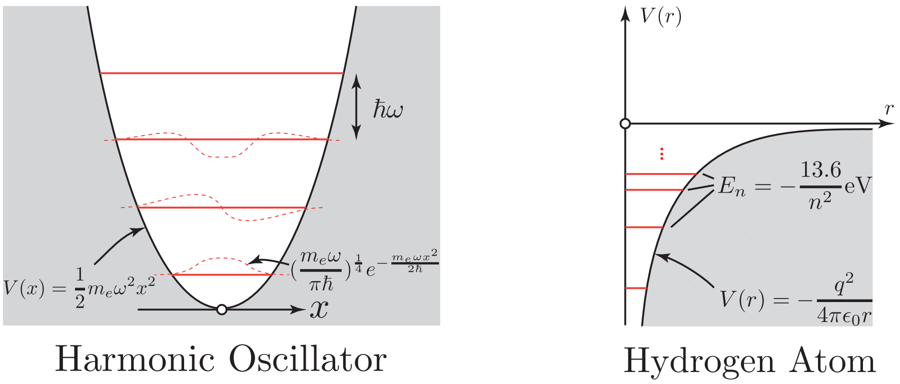
$$\hat{x} = \sqrt{\frac{\hbar}{2m\omega}}(a^\dagger + a) \tag{5.48}$$

$$\hat{p} = i\sqrt{\frac{m\omega\hbar}{2}}(a^\dagger - a) \tag{5.49}$$

$$[a, a^\dagger] = 1 \tag{5.50}$$

$$a|n\rangle = \sqrt{n}|n-1\rangle \tag{5.51}$$

$$a^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle \tag{5.52}$$


$\hbar\omega$
$V(x) = \frac{1}{2}m_e\omega^2x^2$
$(\frac{m_e\omega}{\pi\hbar})^{\frac{1}{4}}e^{-\frac{m_e\omega x^2}{2\hbar}}$
x
$V(r)$
r
$E_n = -\frac{13.6}{n^2}\text{eV}$
$V(r) = -\frac{q^2}{4\pi\epsilon_0 r}$

Harmonic Oscillator

Hydrogen Atom

Fig. 5.8: XX.

5.10 The Hydrogen atom

$$V(r) = -\frac{q^2}{4\pi\epsilon_0 r} \tag{5.53}$$

5.11 Origin of the Elements: The Periodic Table

Closed shells, periodic repeat of active and “noble” gases...

5.12 Origin of the Chemical Bond

Ionic, covalent, van der Waals, hydrogen, etc. Strive to become noble!!

5.13 Electrons in a periodic potential: Bloch Theorem

We finally consider an electron in a periodic potential,

$$-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2}\psi(x)+V(x)\psi(x)=E\psi(x), \tag{5.54}$$

where $V(x+a)=V(x)$. In the absence of the potential $V(x)$, the wavefunctions were of the form $\psi_0(x)=Ae^{ikx}$, where k was allowed to take all values. If we considered a ring of length L, then $k_n=\frac{2\pi}{L}n$, and $\psi(n,x)=\frac{1}{\sqrt{L}}e^{ik_nx}$. Imagine the ring has a periodic lattice, such that $L=Na$. Then, $k_n=\frac{2\pi}{a}\frac{n}{N}$, where $n=0,1,...,N-1$.

← ← **Preceeding text till the right arrows is work in progress**

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 6

Red or Blue pill: Befriending the Matrix

Contents

6.1 The Expansion Principle . 55
6.2 Matrix Mechanics 56
6.3 Matrices and Algebraic Functions 58
6.4 Looking ahead 61
Summary 61
Problems 62

Prior to Schrodinger's differential equation form of 'wave-mechanics' for finding the allowed quantum states of electrons, Heisenberg, Born, and Jordan had developed the first complete form of quantum mechanics, but in the form of matrices. They had named it **Matrix Mechanics**. With the goal to explain the experimentally observed sharp and discrete spectral lines of the Hydrogen atom, Heisenberg hit upon the crucial idea that if the dynamical variables of the electron such as its location $[x]$ and its momentum $[p]$ were *matrices* instead of numbers, then its energy would be a found from a matrix eigenvalue equation, which can yield discrete transition energies. Today we all know that matrices can have discrete eigenvalues, but this connection was not clear in the 1920s when matrices were rarely used in physics. John von Neumann, who was David Hilbert's (Figure 6.1) student, later proved the complete equivalence of Heisenberg's matrix mechanics, and Schrodinger's wave mechanics. In Chapter 5, we became acquainted with the wave-mechanics method of Schrodinger and applied it to the free electron in various dimensions, and a few other problems. In this chapter, we befriend the Matrix method of solving for the quantum mechanical states and energies of electrons. For most numerical solutions, this is the method of choice. With further judicious choice, the matrix equation can give analytical solutions, as we will see in several following chapters for the electron bandstructure in periodic potentials, the situation encountered for semiconductors. We first motivate matrix mechanics by discussing one of the most important and least emphasized principles of quantum mechanics.

Fig. 6.1: David Hilbert, a German mathematician in Gottingen who among many other subjects, developed the idea of Hilbert spaces, which are infinite dimensional matrices with special significance in quantum mechanics. In a delightful story which remains to be confirmed, Hilbert had promised a seminar in the USA on the solution of the famous Fermat's last theorem, to which Fermat had claimed he had a proof but the margin was too small to hold it. The packed audience was disappointed that his seminar had nothing to do with the Fermat's theorem. When asked, Hilbert replied his seminar title was just in case his plane crashed.

6.1 The Expansion Principle

Fourier's theorem mathematically guarantees that *any* well-behaved function $f(x)$ can be expressed as a sum over a *complete* set of trigonometric functions (or complex exponentials): $f(x) = \sum_k a_k e^{ikx}$. Note that *any* complete set of eigenfunctions $[\ \ldots, e^{ikx}, \ldots]$ works! This set has an infinite number of elements and is called the Hilbert space. In practice we typically use a restricted set for most problems. To find the Fourier coefficients, we use the 'filtering' property of complex exponentials $a_{k_n} = \int dx f(x) e^{-ik_n x}$. If we tweak the function $f(x) \rightarrow f(x) + \delta(x) = h(x)$, then $h(x) = \sum_k a^{'}_k e^{ikx}$ is still a valid expansion; the Fourier coefficients will be tweaked from $a_k \rightarrow a^{'}_k$. But note that the perturbed function can still be expanded in terms of the *original* complete set of eigenfunctions. This idea leads to the Expansion Principle in quantum mechanics.

Here is the **Expansion Principle** of quantum mechanics: Any quantum state 'vector' $|\Psi\rangle$ may be expanded as a linear superposition of the eigenvectors of *any* Hermitian operator $|\Psi\rangle = \sum_n a_n |n\rangle$. For most problems, the Hermitian operator of choice is the Hamiltonian operator $\hat{H} = \frac{\hat{\mathbf{p}}^2}{2m_0} + V(\mathbf{r})$, but it need not be. We choose the Hamiltonian operator since there exist a few problems which we encountered in Chapter 5 for which we know the set of *exact* eigenvectors $[\ldots, |n-1\rangle, |n\rangle, |n+1\rangle, \ldots]$. These sets of eigenvectors are *complete*. We also discussed in chapter 5 that this choice of eigenstates are *stationary*. For example, we found the 1D electron on a ring problem with $\hat{H} = \frac{\hat{p}^2}{2m_0}$: gave eigenvalues $E(\mathbf{k}) = \frac{\hbar^2 k^2}{2m_0}$, and corresponding eigenvectors $|k_{\mathrm{FE}}\rangle$ projected to real space $\langle x|k_{\mathrm{FE}}\rangle = \frac{1}{\sqrt{L}} e^{ikx}$. This eigenstate basis $[\ldots, e^{ikx}, \ldots]$ is complete, where k takes all allowed values.

Now consider any state of a harmonic oscillator $|\psi_{\mathrm{HO}}\rangle$. The Expansion Principle guarantees that we can expand *any* harmonic oscillator state in the basis of the free electron $|\psi_{\mathrm{HO}}\rangle = \sum_k a_k |k_{\mathrm{FE}}\rangle$. We can do the reverse too: expand the free electron states in terms of the Harmonic oscillator. This is allowed as long as the potential term in the Hermitian operator does not blow up. For example, we can expand the particle in a box states in terms of the free electron states $|\psi_{box}\rangle = \sum a_k |k_{FE}\rangle$, *but not the other way around* because the particle in a box potential blows up outside the box. This should be obvious because the eigenfunctions of the particle in a box are all ZERO outside the box, and no matter how clever one is, it is not possible to linearly combine zeroes to produce a function that takes non-zero values outside the box.

The Expansion Principle is the backbone of *perturbation theory*, which underpins the quantum mechanics in semiconductor physics. In this chapter, we set up the framework for using it by describing the matrix representation of quantum mechanics.

6.2 Matrix Mechanics

Since we can express any quantum state as an expansion in the eigenvectors $|\Psi\rangle = \sum_n a_n |n\rangle$, we can arrange the expansion coefficients as a column vector

$$|\Psi\rangle = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} \langle 1|\Psi\rangle \\ \langle 2|\Psi\rangle \\ \langle 3|\Psi\rangle \\ \vdots \end{bmatrix}. \tag{6.1}$$

The Hermitian conjugate is obtained by transposing and taking term-by-term complex conjugation:

$$\langle\Psi| = \begin{bmatrix} a_1^\star & a_2^\star & a_3^\star & \cdots \end{bmatrix}, \tag{6.2}$$

which is a row vector. If the state $|\Psi\rangle$ is normalized, then clearly $\langle\Psi|\Psi\rangle = 1$, which requires $\sum_n |a_n|^2 = 1$. Upon measurement, the state will *always* materialize one of the eigenstates $|n\rangle$. Then $|a_n|^2$ should be interpreted as the probability the quantum state materializes in state $|n\rangle$, and a_n is the corresponding probability amplitude. The normalization condition $\sum_n |a_n|^2 = 1$ makes sure that the probabilities add up to one, and the particles are neither created nor destroyed, but their number stay fixed. Also, the state of eigenvectors $|n\rangle$ can always be chosen to be mutually orthogonal, i. e., $\langle m|n\rangle = \delta_{mn}$. This is the basis we will work with.

Fig. 6.2: Pascual Jordan, who with Heisenberg and Max Born in Gottingen developed the first consistent version of quantum mechanics in its *matrix* form, and named it Matrix Mechanics. Jordan also did seminal work on quantum field theory. Had he not associated himself with the Nazi party, he would have been recognized today as well as Heisenberg and Born.

Also note that projecting $|\Psi\rangle = \sum_n a_n |n\rangle$ on state $\langle n|$ yields the coefficients $a_n = \langle n|\Psi\rangle$. Then, we can write the expansion as $|\Psi\rangle = \sum_n |n\rangle\langle n|\Psi\rangle$, which means that

$$\sum_n |n\rangle\langle n| = 1. \tag{6.3}$$

This is the 'closure' relation of eigenvectors that are discrete. If the eigenvectors were continuous, then the corresponding closure relation is

$$\int dx |x\rangle\langle x| = 1. \tag{6.4}$$

The fact that the two closure relations are unity allows us to insert them wherever they will help in the evaluation of matrix elements. Consider now an operator $\hat{A}$ acting on the state vector $|\Psi\rangle$. It will try to 'rotate' the state vector in the Hilbert space to a state $|\Phi\rangle$ as shown pictorially in Figure 6.3 . We write this as

$$\hat{A}|\Psi\rangle = |\Phi\rangle. \tag{6.5}$$

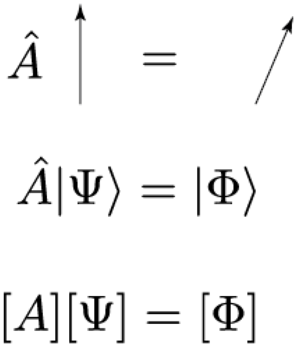


Fig. 6.3: Three ways of saying the same thing. The operator $\hat{A}$ rotates a state vector $|\Psi\rangle$ into $|\Phi\rangle$. The pictorial depiction is equivalent to the algebraic operator equation, which in turn is equivalent to the matrix form $[A][\Psi] = [\Phi]$.

By the expansion principle, we can expand the new state $|\Phi\rangle = \sum_m b_m|m\rangle$. Then, if we project this state on $|m\rangle$, we have

$$\langle m|\Phi\rangle = \langle m|\hat{A}|\Psi\rangle \to b_m = \sum_n a_n\langle m|\hat{A}|n\rangle = \sum_n A_{mn}a_n. \tag{6.6}$$

We see that the operator is equivalent to a *matrix* $\hat{A} \equiv A_{mn} = [A]$. The elements of the equivalent matrix are the terms $A_{mn} = \langle m|\hat{A}|n\rangle$, obtained by the operator acting on eigenstates on both sides. We call them **matrix elements** for obvious reasons.

For example, if we choose the momentum operator acting between states $|k\rangle$ and $\langle k'|$ of the free electron, we get $p_{kk'} = \langle k'|\hat{p}|k\rangle = \int dx\langle k'|x\rangle\langle x|\hat{p}|k\rangle = \hbar k\delta_{k',k}$. Note that the 'abstract' operator $\hat{p}$ has the matrix representation $\langle x|\hat{p}|x\rangle = -i\hbar\frac{\partial}{\partial x}$ in real space. The example shows that since the free-electron energy eigenstates are simultaneously momentum eigenstates, the momentum operator acting between two eigenstates extracts the value of the momentum only if the two states are identical. This is the momentum matrix element.

One of the most important operators is the Hamiltonian operator, which 'extracts' the energy of the state it is acting on. If the state $|n\rangle$ happens to be an eigenstate, the Hamiltonian operator extracts its energy eigenvalue: $\hat{H}|n\rangle = E_n|n\rangle$. Visualize $\hat{H}|n\rangle$ as a new vector whose 'direction' is the same as the eigenvector $|n\rangle$, but the length determined by the eigenvalue E_n. So the action of the Hamiltonian operator leaves the 'direction' of the eigenvector $|n\rangle$ unaffected.

If the state is not an energy eigenstate but is a linear superposition $|\Psi\rangle = \sum_n a_n|n\rangle$, then the time-independent Schrodinger equation states that $\hat{H}|\Psi\rangle = E|\Psi\rangle$, which is equivalent to $\hat{H}\sum_n a_n|n\rangle = E\sum_n a_n|n\rangle$. When we project this new state vector $\hat{H}|\Psi\rangle$ on the eigenvector $\langle m|$, we get an algebraic equation

$$\sum_n \langle m|\hat{H}|n\rangle a_n = Ea_m, \tag{6.7}$$

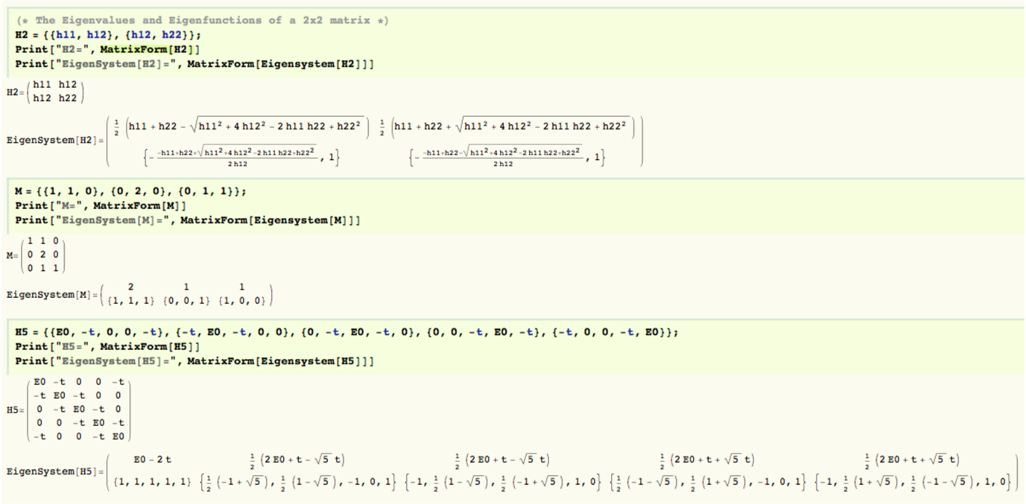
for each m. Note the appearance of the matrix elements $H_{mn} = \langle m|\hat{H}|n\rangle$. If we write this out for $m = 1, 2, \ldots,$ we get the set of linear equations

$$\begin{array}{lcl} H_{11}a_1 + H_{12}a_2 + H_{13}a_3 \ldots & = & Ea_1 \\ H_{21}a_1 + H_{22}a_2 + H_{23}a_3 \ldots & = & Ea_2 \\ H_{31}a_1 + H_{32}a_2 + H_{33}a_3 \ldots & = & Ea_3 \\ \vdots & = & \vdots \end{array} \tag{6.8}$$

which is best captured as a matrix equation

$$\begin{bmatrix} H_{11} & H_{12} & H_{13} & \ldots \\ H_{21} & H_{22} & H_{23} & \ldots \\ H_{31} & H_{32} & H_{33} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = E \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix}. \tag{6.9}$$

Note that the Hamiltonian operator becomes a square matrix, and the state $|\Psi\rangle$ becomes a column vector. This matrix equation contains the same information as the algebraic time-independent Schrodinger equation. If we choose to work with a restricted set of say 10 states, then we have a 10x10 matrix with 10 eigenvalues and their corresponding eigenfunctions. Figure 6.4 shows a few examples of matrix evaluations using the Mathematica package.



```
(* The Eigenvalues and Eigenfunctions of a 2x2 matrix *)
H2 = {{h11, h12}, {h12, h22}};
Print["H2=", MatrixForm[H2]]
Print["EigenSystem[H2]=", MatrixForm[Eigensystem[H2]]]
```

$$\text{H2=}\begin{pmatrix} \text{h11} & \text{h12} \\ \text{h12} & \text{h22} \end{pmatrix}$$

$$\text{EigenSystem[H2]=}\begin{pmatrix} \frac{1}{2}\left(\text{h11}+\text{h22}-\sqrt{\text{h11}^2+4\,\text{h12}^2-2\,\text{h11 h22}+\text{h22}^2}\right) & \frac{1}{2}\left(\text{h11}+\text{h22}+\sqrt{\text{h11}^2+4\,\text{h12}^2-2\,\text{h11 h22}+\text{h22}^2}\right) \\ \left\{-\frac{-\text{h11}+\text{h22}+\sqrt{\text{h11}^2+4\,\text{h12}^2-2\,\text{h11 h22}+\text{h22}^2}}{2\,\text{h12}},1\right\} & \left\{-\frac{-\text{h11}+\text{h22}-\sqrt{\text{h11}^2+4\,\text{h12}^2-2\,\text{h11 h22}+\text{h22}^2}}{2\,\text{h12}},1\right\} \end{pmatrix}$$

```
M = {{1, 1, 0}, {0, 2, 0}, {0, 1, 1}};
Print["M=", MatrixForm[M]]
Print["EigenSystem[M]=", MatrixForm[Eigensystem[M]]]
```

$$\text{M=}\begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

$$\text{EigenSystem[M]=}\begin{pmatrix} 2 & 1 & 1 \\ \{1,1,1\} & \{0,0,1\} & \{1,0,0\} \end{pmatrix}$$

```
H5 = {{E0, -t, 0, 0, -t}, {-t, E0, -t, 0, 0}, {0, -t, E0, -t, 0}, {0, 0, -t, E0, -t}, {-t, 0, 0, -t, E0}};
Print["H5=", MatrixForm[H5]]
Print["EigenSystem[H5]=", MatrixForm[Eigensystem[H5]]]
```

$$\text{H5=}\begin{pmatrix} \text{E0} & -\text{t} & 0 & 0 & -\text{t} \\ -\text{t} & \text{E0} & -\text{t} & 0 & 0 \\ 0 & -\text{t} & \text{E0} & -\text{t} & 0 \\ 0 & 0 & -\text{t} & \text{E0} & -\text{t} \\ -\text{t} & 0 & 0 & -\text{t} & \text{E0} \end{pmatrix}$$

$$\text{EigenSystem[H5]=}\begin{pmatrix} \text{E0}-2\,\text{t} & \frac{1}{2}\left(2\,\text{E0}+\text{t}-\sqrt{5}\,\text{t}\right) & \frac{1}{2}\left(2\,\text{E0}+\text{t}-\sqrt{5}\,\text{t}\right) & \frac{1}{2}\left(2\,\text{E0}+\text{t}+\sqrt{5}\,\text{t}\right) & \frac{1}{2}\left(2\,\text{E0}+\text{t}+\sqrt{5}\,\text{t}\right) \\ \{1,1,1,1,1\} & \left\{\frac{1}{2}\left(-1+\sqrt{5}\right),\frac{1}{2}\left(1-\sqrt{5}\right),-1,0,1\right\} & \left\{-1,\frac{1}{2}\left(1-\sqrt{5}\right),\frac{1}{2}\left(-1+\sqrt{5}\right),1,0\right\} & \left\{\frac{1}{2}\left(-1-\sqrt{5}\right),\frac{1}{2}\left(1+\sqrt{5}\right),-1,0,1\right\} & \left\{-1,\frac{1}{2}\left(1+\sqrt{5}\right),\frac{1}{2}\left(-1-\sqrt{5}\right),1,0\right\} \end{pmatrix}$$

Fig. 6.4: Examples of 2x2, 3x3, and 5x5 Matrix eigenvalue and eigenfunction calculations in Mathematica. The 2x2 Hamiltonian is general and one of the most important in all of quantum mechanics. The 3x3 matrix is a numerical example, and the 5x5 matrix of a 5-site circular ring tight-binding Hamiltonian model. Note that the eigenvectors (or eigenfunction coefficients a_n are evaluated for each eigenvalue, which is *very* nice.

Indeed, historically Heisenberg developed quantum mechanics in its matrix representation and called it 'matrix mechanics'. Schrodinger found the algebraic version which appealed more to researchers since we are trained much better in algebra since high school than in matrices. But they are one and the same thing.

6.3 Matrices and Algebraic Functions

Numbers are solutions to algebraic equations. We start our education learning about integers because they quantify the fingers on our hand, and soon expand into the regime of real numbers. Soon, we realize there are algebraic equations such as $x^2 + 1 = 0$ which have solutions that are not real numbers, and realize there must be new kinds of numbers. Complex numbers contain $i = \sqrt{-1}$, which (unfortunately[1]) is called an imaginary number. We learn how to add, subtract, multiply, divide, take square roots, exponentiate, etc... with numbers.

One can visualize a matrix as an extension of the concept of a 'number'. For example, the algebraic equation $ax = b$ has the solution $x = b/a$, a number. A *set* of algebraic equations with multiple variables can be written in the form $AX = B$ has the solution $X = A^{-1}B$. Somewhere along the line, if we do not use matrices, we forget their power and beauty! We get busy using algebraic equations extensively. Turns out every algebraic equation may be written as a matrix equation. Then we can use powerful theorems of matrices to solve or analyze them. Indeed, most *numerical* approaches to solving equations have to go through the matrix route. Consider the equation of a unit circle

$$x^2 + y^2 = 1. \tag{6.10}$$

[1] One might think ... that imaginary numbers are just a mathematical game having nothing to do with the real world. From the viewpoint of positivist philosophy, however, one cannot determine what is real. All one can do is find which mathematical models describe the universe we live in. It turns out that a mathematical model involving imaginary time predicts not only effects we have already observed but also effects we have not been able to measure yet nevertheless believe in for other reasons. So what is real and what is imaginary? Is the distinction just in our minds? - Stephen Hawking

This may not look like a matrix equation at all, till we define the coordinate 'vector' $X = \begin{bmatrix} x \\ y \end{bmatrix}$. Its transpose is a row vector $X^T = [x, y]$, and the matrix version of the equation of the unit circle is then

$$X^T X = 1. \tag{6.11}$$

Now consider the equation

$$x^2 + axy + y^2 = 1, \tag{6.12}$$

which can be written as

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & u \\ a-u & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = X^T A X = 1. \tag{6.13}$$

This works for any value of u. So for the unit circle, $A = I$, where I is the unit matrix. The matrix A captures the *geometric* property of the curve - whether it is a circle, or a more complex shape. Later in this book we will investigate how such decompositions actually help understand the *curvature* of a function, where A will take the form of a Hessian matrix, and find the 'curvature' of the allowed eigenvalues of a quantum problem.

The strangest and most striking property of matrices is that they do not necessarily *commute*. Which is to say that in general for square matrices, $AB \neq BA$. As a mathematical object, therefore they are quite distinct from real or complex numbers. Matrices thus form the natural objects for non-commutative algebra. Therefore they are central to the tenets of quantum mechanics, which is built upon the non-commutativity of operators as embodied by $\hat{x}\hat{p}_x - \hat{p}_x\hat{x} = i\hbar$, which actually was derived by Heisenberg and Born for the first time in its matrix form $[x][p] - [p][x] = i\hbar[I]$.

A square matrix A has eigenvalues λ_i and eigenvectors $[x_i]$ which are obtained by solving the equation

$$A[x] = \lambda[x] \rightarrow [A - \lambda I][x] = 0 \tag{6.14}$$

After finding the eigenvalues and eigenvectors, the square matrix can be re-written in it's 'spectral' decomposition

$$A = UDU^{-1}, \tag{6.15}$$

where D is the diagonal matrix

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots \\ 0 & \lambda_2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{6.16}$$

and the unitary transformation matrix U is formed by arranging the eigenvectors in the same order as the eigenvalues

$$U = \begin{bmatrix} [x_1] & [x_2] & \cdots \end{bmatrix} \tag{6.17}$$

Note that U is invertible, meaning its determinant cannot be zero.

Now lets say the square matrix A is actually the Hamiltonian matrix of a quantum mechanics problem. Then solving the time-independent Schrodinger equation is *equivalent* to diagonalizing the Hamiltonian matrix by solving the matrix equation

$$\begin{bmatrix} H_{11} & H_{12} & H_{13} & \dots \\ H_{21} & H_{22} & H_{23} & \dots \\ H_{31} & H_{32} & H_{33} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = E \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix}. \tag{6.18}$$

Clearly, it is equivalent to

$$\begin{bmatrix} H_{11}-E & H_{12} & H_{13} & \dots \\ H_{21} & H_{22}-E & H_{23} & \dots \\ H_{31} & H_{32} & H_{33}-E & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = 0. \tag{6.19}$$

If instead of the infinite matrix, we choose a restricted eigenbasis set of N, then the solutions of the corresponding algebraic equation $\mathrm{Det}[H - EI] = 0$ yield N eigenvalues E_n. Corresponding to each eigenvalue E_n, there is an eigenvector $|n\rangle$ which is a column vector. We then construct the unitary operator U by collecting the eigenvectors and write the Hamiltonian matrix in its diagonal form as

$$\begin{bmatrix} H_{11} & H_{12} & H_{13} & \dots & H_{1N} \\ H_{21} & H_{22} & H_{23} & \dots & H_{2N} \\ H_{31} & H_{32} & H_{33} & \dots & H_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ H_{N1} & H_{N2} & H_{N3} & \dots & H_{NN} \end{bmatrix} = U \begin{bmatrix} E_1 & 0 & 0 & \dots & 0 \\ 0 & E_2 & 0 & \dots & 0 \\ 0 & 0 & E_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & E_N \end{bmatrix} U^{-1} \tag{6.20}$$

This is the 'spectral' decomposition of the Hamiltonian $H = UDU^{-1}$ where D is a diagonal matrix whose elements are the energy eigenvalues. The exact solution requires the matrices to be infinite-dimensional, but for most practical cases we work with a restricted set.

Now lets imagine that the Hamiltonian matrix is perturbed to $H \to H_0 + W$. The eigenvalues and eigenfunctions will change. But the expansion principle tells us that the new state vector of the perturbed system can still be expanded in terms of the unperturbed eigenvectors, or the matrix U. The matrix formalism makes such perturbations easy to deal with. We will return to this problem in the next chapter.

The sum of the diagonal elements of a square matrix is called its trace, $\mathrm{Tr}[H] = \sum_n H_{nn}$. For square matrices A, B, $\mathrm{Tr}[AB] = \mathrm{Tr}[BA]$. Thus, we get $\mathrm{Tr}[H] = \mathrm{Tr}[UDU^{-1}] = \mathrm{Tr}[U^{-1}UD] = \mathrm{Tr}[D] = \sum_n E_n$. The trace of the Hamiltonian is the sum of its eigenvalues.

The quantum states are represented as column vectors $|\Psi\rangle = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix}$, as discussed earlier. Consider another quantum state $|\Phi\rangle = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \end{bmatrix}$. If we take the projection $\langle\Phi|\Psi\rangle = \sum_n a_n b_n^\star$, we get a *number*. This is analogous to taking the dot product of two vectors, and is called the 'inner' product for Dirac notation. But we can also take an 'outer' product

$$|\Psi\rangle\langle\Phi| = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} \begin{bmatrix} b_1^\star & b_2^\star & \dots & b_N^\star \end{bmatrix} = \begin{bmatrix} a_1b_1^\star & a_1b_2^\star & a_1b_3^\star & \dots & a_1b_N^\star \\ a_2b_1^\star & a_2b_2^\star & a_2b_3^\star & \dots & a_2b_N^\star \\ a_3b_1^\star & a_3b_2^\star & a_3b_3^\star & \dots & a_3b_N^\star \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_Nb_1^\star & a_Nb_2^\star & a_Nb_3^\star & \dots & a_Nb_N^\star \end{bmatrix}, \tag{6.21}$$

which is no longer a number but a *matrix*. This matrix clearly contains information of the phases of each quantum state and their respective projections on eigenstates, which is lost in taking the inner product. The outer product leads to the concept of density matrices, which keep track of the phase relationships and interferences of quantum states. An interesting result is that the trace of the inner and outer products are the same, i.e.

$\text{Tr}[|\Psi\rangle\langle\Phi|] = \text{Tr}[\langle\Phi|\Psi\rangle] = \sum_n a_n b_n^\star$. The mathematical name for objects like the outer product $|m\rangle\langle n|$ is a *dyadic*, it is a tensor constructed out of two vectors.

That the outer product is a matrix implies we can think of it as an *operator*. In fact, we can construct operators of the form $\hat{A} = \sum a_n |n\rangle\langle n|$ with suitable coefficients a_n. One such construction will prove rather useful. Consider the Schrodinger equation $\hat{H}|n\rangle = E_n|n\rangle$ with eigenvalues E_n and eigenvectors $|n\rangle$. We define a new operator

$$\hat{G}(E) = \sum_n \frac{|n\rangle\langle n|}{E - E_n}, \tag{6.22}$$

where the coefficients are $a_n = 1/(E - E_n)$, with units of inverse energy. This operator is called the Green's function operator. We will use it in the later chapters. For now, note that it blows up every time $E = E_n$, and changes sign as E crosses E_n. Note what happens when the Green's function operator acts on an eigenstate $|m\rangle$:

$$\hat{G}(E)|m\rangle = \sum_n \frac{|n\rangle\langle n|}{E - E_n}|m\rangle = \sum_n \frac{|n\rangle}{E - E_n}\langle n|m\rangle = \frac{1}{E - E_m}|m\rangle. \tag{6.23}$$

This happens for every eigenvalue, because $|m\rangle$ can be any of the eigenvectors. We can picturize the Green's function operator as a 'energy-eigenvalue detector'. As we sweep the energies, every time $E = E_m$, there is a very large response since $\hat{G}(E_m)|m\rangle \to \pm\infty$. The response is low between eigenvalues is low. This is analogous to the concept of a 'impulse response' in linear systems.

The Schrodinger equation may be written as $(E - \hat{H}^0)|\psi\rangle = 0$. Then note that if we apply the Green's function operator on the left side of the equation, we get

$$\hat{G}(E)(E - \hat{H}^0)|\psi\rangle = \sum_n \frac{|n\rangle\langle n|}{E - E_n}(E - \hat{H}^0)|\psi\rangle = \sum_n |n\rangle\langle n|\psi\rangle = |\psi\rangle. \tag{6.24}$$

From the above, it is clear that $\hat{G}(E) = (E - \hat{H}^0)^{-1}$, i.e., the Green's function operator is the inverse operator of $(E - \hat{H}^0)$. You can think of this in terms of matrices to make it more concrete. Also, the right side of the Schrodinger equation was zero, meaning $\hat{G}(E)0 = |\psi\rangle$. This may seem weird because the Green's function operator seems to act on 'zero' to create the state $|\psi\rangle$. We will return to this strange behavior in chapter **??** to explore what it is trying to say.

6.4 Looking ahead

What is the advantage of the spectral decomposition of a matrix $A = UDU^{-1}$? Let's observe what happens when we try to square the matrix A.

$$A^2 = UD(U^{-1}U)DU^{-1} = UD^2U^{-1} = U \begin{bmatrix} \lambda_1^2 & 0 & \cdots \\ 0 & \lambda_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} U^{-1}. \tag{6.25}$$

$U^{-1}U = I$ contracts the expansion, and we only have to square the diagonal matrix, which is trivial since we just square the eigenvalues! Think of any higher order power of the matrix: the U's always contract, so $A^N = UD^NU^{-1}$! If we visualize matrices as extensions of real and complex numbers, we should be curious about doing similar operations on them. For example, what is the square root of a matrix? What is the exponential or logarithm of a matrix? Can we take sines and cosines of matrices? The answer to all of these questions is yes, and the spectral decomposition is the first step to such fun. In this chapter, we discussed the *linear* properties of matrices, which will help us get started with perturbation theory.

Chapter Summary

.. . .

Problems

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 7

Perturbations to the Electron's Freedom

Contents

7.1 Degenerate Perturbation Theory 63
7.2 Non-Degenerate Perturbation Theory 65
7.3 The Brillouin-Wigner Perturbation Results . 68
7.4 Rayleigh-Schrodinger Perturbation Results . . . 68
7.5 The Hellmann Feynman Theorem 70
7.6 Perturbation Theory Example 70
Summary 72
Problems 72

In chapter 5 we discussed a few *exactly* solved problems in quantum mechanics. We also found that the many applied problems may not be exactly solvable in an analytical form. The machinery to solve such problems is called perturbation theory. In chapter 6, we developed the matrix formalism of quantum mechanics, which is well-suited to handle perturbation theory. Sometimes we will be able to reduce the matrix solutions to closed-form algebraic forms which always helps in visualization. in this chapter, we develop an additional analytical tool for perturbation theory which is indispensable in the insight it provides.

Let $\hat{H}^0$ be the Hamiltonian for the solved problem. Then the time-dependent Schrodinger equation is $i\hbar\frac{\partial}{\partial t}|\Psi\rangle = \hat{H}^0|\Psi\rangle$. The eigenstates of definite energy are also *stationary* states $\langle r|n\rangle = \psi_E(r)e^{-iE_n t/\hbar}$, where $|n\rangle$ are the eigenvectors and E_n the corresponding eigenvalues. Note that *all* the solved problems we discussed in chapter 5 such as the harmonic oscillator or the particle in a box had *time-independent* potentials. Many real-world situations involve time-dependent potentials. For example, imagine a field-effect transistor whose gate voltage is being modulated by an ac signal. That will create a potential variation for electrons of the form $V(r)e^{i\omega t}$. A similar variation will be experienced by electrons interacting with photons of an electromagnetic wave, or with phonons of lattice vibrations. Consider the limit of very small frequencies $\omega \to 0$, or a 'dc' potential. Then, the potential only has a spatial variation. A dc voltage is not truly time-independent because it has to be turned on or off at some time. But most of the physics we are interested in this and a few following chapters happens when the perturbation has been 'on' for a ling time in the past, and things have reached steady-state. It is in this sense that we discuss time-independent perturbation theory. We defer explicitly time-varying or oscillatory perturbations to later chapters.

7.1 Degenerate Perturbation Theory

The time-independent Schrodinger equation for the *solved* problem is

$$\hat{H}^0|n\rangle = E_n^0|n\rangle, \tag{7.1}$$

where $\hat{H}^0$ is the unperturbed Hamiltonian. That means we know all the eigenfunctions $|n\rangle$ and their corresponding eigenvalues E_n^0. This is shown in Fig 7.1. Lets add a perturbation W to the initial Hamiltonian such that the new Hamiltonian becomes $\hat{H} = \hat{H}^0 + W$. The new Schrodinger equation is

$$(\hat{H}^0 + W)|\psi\rangle = E|\psi\rangle. \tag{7.2}$$

The perturbation W has changed the eigenvectors $|n\rangle \rightarrow |\psi\rangle$. The corresponding eigenvalues may not be eigenvalues of the new Hamiltonian. Some eigenvalues increase in energy, some decrease, and others may not be affected. This is illustrated in Fig 7.1. So we have to solve for the new eigenvalues E and obtain the corresponding eigenvectors.

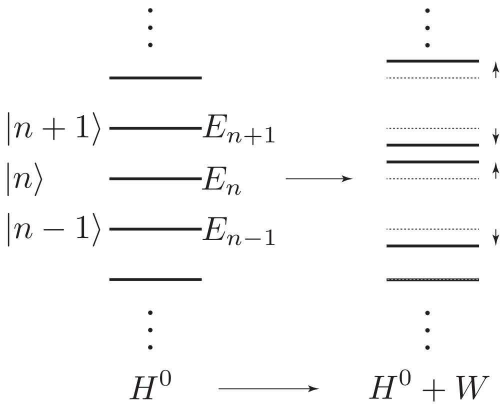


Fig. 7.1: The initial eigenstates and eigenvalues of a quantum system change upon application of a perturbation W.

At this stage, we invoke the **Expansion Principle** introduced in chapter 6. It states that the *perturbed* state vector $|\psi\rangle$ can always be written as a linear superposition of the *unperturbed* eigenvectors $|n\rangle$, since the unperturbed eigenstates form a complete basis set. It is the same philosophy of expanding any function in terms of its Fourier components. Thus we write

$$|\psi\rangle = \sum_n a_n |n\rangle, \tag{7.3}$$

where a_n's are (in general complex) expansion coefficients. The coefficients are obtained by taking the projection $\langle m|\psi\rangle$, which yields $a_n = \langle n|\psi\rangle$. Then equation 7.2 reads

$$\sum_n a_n(\hat{H}^0 + W)|n\rangle = E\sum_n a_n|n\rangle. \tag{7.4}$$

We can visualize the new state vector as the original eigenvector 'rotated' by the perturbation W, as we had introduced in Chapter 6, and specifically in Figure 6.3. Lets project the new state vector on $\langle m|$ to get

$$\sum_n a_n\langle m|(\hat{H}^0 + W)|n\rangle = Ea_m, \tag{7.5}$$

which is a matrix when m takes values $1, 2, \ldots N$

$$\begin{bmatrix} E_1+W_{11} & W_{12} & W_{13} & \ldots & W_{1N} \\ W_{21} & E_2+W_{22} & W_{23} & \ldots & W_{2N} \\ W_{31} & W_{32} & E_3+W_{33} & \ldots & W_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_{N1} & W_{N2} & W_{N3} & \ldots & E_N+W_{NN} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_N \end{bmatrix} = E \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_N \end{bmatrix}. \tag{7.6}$$

The eigenvalues and the corresponding eigenvectors of this matrix equation are obtained by diagonalization, as discussed in chapter 6. The new eigenvalues $E_n^{'}$ thus depend on the matrix elements $W_{mn} = \langle m|W|n\rangle$ of the perturbation. Note that if some eigenvalues of the unperturbed Hamiltonian happened to be degenerate, the matrix diagonalization method takes that into account naturally without problems. In that sense, the matrix formulation of perturbation theory is sometimes referred to as *degenerate* perturbation theory. But the matrix formulation handles non-degenerate situations equally well, and is more general.

In case we did not start with a diagonal basis of the unperturbed Hamiltonian H^0, then we have the Schrodinger equation

$$\begin{bmatrix} H^0_{11}+W_{11} & H^0_{12}+W_{12} & H^0_{13}+W_{13} & \dots & H^0_{1N}+W_{1N} \\ H^0_{21}+W_{21} & H^0_{22}+W_{22} & H^0_{23}+W_{23} & \dots & H^0_{2N}+W_{2N} \\ H^0_{31}+W_{31} & H^0_{32}+W_{32} & H^0_{33}+W_{33} & \dots & H^0_{3N}+W_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ H^0_{N1}+W_{N1} & H^0_{N2}+W_{N2} & H^0_{N3}+W_{N3} & \dots & H^0_{NN}+W_{NN} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_N \end{bmatrix} = E \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_N \end{bmatrix}. \tag{7.7}$$

The solutions thus reduce to diagonalizing the corresponding perturbed matrices. Many of the perturbation matrix elements $W_{mn} = \langle m|W|n\rangle$ can be made zero by suitable choice of bases, which reduces the work involved in diagonalization. Note that we can easily obtain the eigenvalues, the eigenvectors typically require more work. But with the availability of math packages such as Mathematica and MATLAB, this is done in a jiffy for most situations we will deal with.

An important question in applying degenerate perturbation theory is: which states should be included in the $N \times N$ matrix? At this stage, we state the guiding principles, we will see the proof of the principle in the next section on non-degenerate perturbation theory. The first principle is that states with eigenvalues widely separated in energy $E_n - E_m$ interact weakly by the perturbation. The second principle is states get pushed around by the perturbation depending upon the matrix element squared $|W_{mn}|^2$. Quantitatively, the perturbation in energy of a state with energy E is $\Delta E \approx |W_{mn}|^2/(E - E_n)$ by interacting with state $|n\rangle$. Thus, states for which W_{mn} terms are small or zero may be left out. If we are interested in a set of energy eigenvalues (say near the conduction and valence band edges of a semiconductor), energies far away from the band edges may also be left out. We will see the application of these rules in many following chapters.

We will see examples of degenerate perturbation theory in the next chapter (chapter 8), where we will apply it to the problem of a free electron. That will require us to solve either 2×2 matrices, or higher order, depending on the accuracy we need. Later on, we will also encounter it when we discuss the $\mathbf{k} \cdot \mathbf{p}$ theory of bandstructure to deal with the degeneracies of heavy and light valence bands. For now, we look at particular situations when we have 'isolated' eigenvalues that are non-degenerate.

7.2 Non-Degenerate Perturbation Theory

Schrodinger's crowning achievement was to obtain an algebraic equation, which when solved, yields the quantum states allowed for electrons. Schrodinger's equation is called the 'wave'-equation because it was constructed in analogy to Maxwell's equations for electromagnetic waves. Heisenberg was the first to achieve the breakthrough in quantum mechanics before Schrodinger, except his version involved matrices. Which is why he called it matrix-mechanics. That is why it was not as readily accepted - again because matrices are unfamiliar to most. It was later that the mathematician von Neumann proved that both approaches were actually identical from a mathematical point of view.

So at this point, we will try to return to a 'familiar' territory in perturbation theory from the matrix version presented in the previous section. We try to formulate an *algebraic* method to find the perturbed eigenvalues and eigenvectors.

Consider a perturbation $\hat{W}$ added to the solved (or unperturbed) Hamiltonian $\hat{H}^0$. Schrodinger equation is

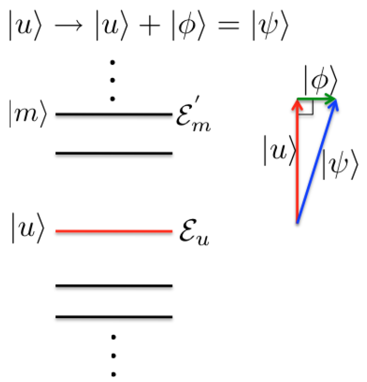


Fig. 7.2: The perturbation rotates the eigenvector $|u\rangle$ to $|\psi\rangle$. If we forego normalization of $|\psi\rangle$, we can find a vector $|\phi\rangle$ orthogonal to $|u\rangle$ such that $\langle u|\phi\rangle = 0$, and consequently $\langle u|\psi\rangle = 1$.

$$(\hat{H}^0 + \hat{W}))|\psi\rangle = E|\psi\rangle, \tag{7.8}$$

and the unperturbed state $|u\rangle$ satisfied

$$\hat{H}^0|u\rangle = E_u|u\rangle. \tag{7.9}$$

The new quantum state differs from the unperturbed state, so we write

$$|\psi\rangle = |u\rangle + |\phi\rangle. \tag{7.10}$$

We can picturize the final state $|\psi\rangle$ as a 'vector' sum of the unperturbed state $|u\rangle$ and the vector $|\phi\rangle$. This is schematically shown in Fig 7.2. In particular, if we are willing to not normalize the final state, then we can always choose $|\phi\rangle$ to be orthogonal to $|u\rangle$, leading to $\langle u|\phi\rangle = 0$ and $\langle u|\psi\rangle = 1$. We can then project equation 7.8 on $\langle u|$ to obtain the energy equation

$$E = E_u + \langle u|W|\psi\rangle = \underbrace{E_u}_{\text{unperturbed}} + \underbrace{\langle u|W|u\rangle}_{\Delta E^{(1)}} + \underbrace{\langle u|W|\phi\rangle}_{\text{higher orders}}. \tag{7.11}$$

Note that we obtain the 'first-order' energy correction: they are the diagonal matrix elements of the perturbation with the unperturbed states. Think of a 'dc' perturbation - say a voltage V_0 that depends neither on space nor time - then all the initial energy eigenvalues get shifted by the corresponding energy: $E_u \rightarrow E_u + qV_0$ due to the first order term since $\langle u|qV_0|u\rangle = qV_0$. We will shortly see that for this particular perturbation, the higher order terms are zero because they depend on the *cross*-matrix terms of the kind $\langle m|qV_0|n\rangle = qV_0\langle m|n\rangle = 0$. An example of such a situation is when a voltage is applied across a gate capacitor to a semiconductor - the entire bandstructure which consists of the allowed E_n's shift rigidly up or down. We call this energy band-bending in device physics. Such a 'dc' perturbation does not couple different energy states for the above reason, and results in only a first-order rigid shift.

Most perturbations are not the 'dc'-kind, and we need the higher order terms for them. To do that, it is useful to define

$$E_u' = E_u + \langle u|W|u\rangle. \tag{7.12}$$

We then split the diagonal and off-diagonal elements of the perturbation just like writing a signal as a 'dc' + a 'ac' terms. Think of $\hat{W}$ as an operator, and hence a matrix that we are splitting it into two:

$$\hat{W} = \hat{D} + \hat{W}', \tag{7.13}$$

and the total Hamiltonian then is

$$\hat{H} = \underbrace{\hat{H}^0 + \hat{D}}_{\hat{H}^{(d)}} + \hat{W}', \tag{7.14}$$

The reason for doing this is that the unperturbed eigenvalues are going to shift by the diagonal part of the perturbation without interacting with other states. The off-diagonal terms will further tweak them by interactions with other states. To move further, we write

$$(\hat{H}^{(d)} + \hat{W}')|\psi\rangle = E|\psi\rangle, \tag{7.15}$$

and rearrange it to

$$(E - \hat{H}^{(d)})|\psi\rangle = \hat{W}'|\psi\rangle, \tag{7.16}$$

At this stage, our goal is to find the perturbation vector $|\phi\rangle = |\psi\rangle - |u\rangle$. How can we obtain it from the left side of equation 7.16 in terms of the perturbation on the right? Recall in chapter **??** we discussed the Green's function operator briefly. We noticed that it is an 'inverse' operator, meaning we expect

$$\hat{G}(E)(E - \hat{H}^{(d)})|\psi\rangle = \sum_m \frac{|m\rangle\langle m|}{E - E'_m}(E - \hat{H}^{(d)})|\psi\rangle = \sum_m |m\rangle\langle m|\psi\rangle = |\psi\rangle. \tag{7.17}$$

So to get $|\phi\rangle = |\psi\rangle - |u\rangle$, perhaps we should use the operator

$$\hat{G}(E) - \frac{|u\rangle\langle u|}{E - E'_u} = \sum_{m \neq u} \frac{|m\rangle\langle m|}{E - E'_m}. \tag{7.18}$$

Operating on the LHS of equation 7.16 we obtain

$$\sum_{m \neq u} \frac{|m\rangle\langle m|}{E - E'_m}(E - \hat{H}^{(d)})|\psi\rangle = \sum_{m \neq u} |m\rangle\langle m|\psi\rangle = (\sum_m |m\rangle\langle m|\psi\rangle) - |u\rangle\langle u|\psi\rangle = |\psi\rangle - |u\rangle = |\phi\rangle, \tag{7.19}$$

which is what we wanted. Now we use the same operator on the right of equation 7.16 to finish the job. Since $\hat{W}'$ consists of only off-diagonal cross matrix elements, we write it in its outer product form as $\hat{W}' = \sum_m \sum_{m \neq n} |m\rangle\langle m|\hat{W}|n\rangle\langle n|$, and apply the 'reduced' Green's function to get

$$|\phi\rangle = \sum_{l \neq u} \sum_m \sum_{n \neq m} \frac{|l\rangle\langle l|}{E - E'_l} |m\rangle\langle m|\hat{W}|n\rangle\langle n|\psi\rangle = \sum_{m \neq u} \sum_{n \neq m} |m\rangle \frac{\langle m|\hat{W}|n\rangle}{E - E'_m}\langle n|\psi\rangle, \tag{7.20}$$

Thus, we obtain the perturbed state $|\psi\rangle = |u\rangle + |\phi\rangle$ to be

$$|\psi\rangle = |u\rangle + \underbrace{\sum_{m \neq u} \sum_{n \neq m} |m\rangle \frac{\langle m|\hat{W}|n\rangle}{E - E'_m}\langle n|\psi\rangle}_{|\phi\rangle}. \tag{7.21}$$

As a sanity check, we note that if $\hat{W} = 0$, $|\psi\rangle = |u\rangle$, as it should be. Next, we note that this is a *recursive* relation, meaning $|\psi\rangle$ also appears inside the sum on the right side. Thus, it can be taken to many orders, but we are going to retain just up to the 2nd order. That means, we will assume that the perturbation is weak, and so we are justified in replacing the $|\psi\rangle$ inside the sum on the right side by the unperturbed state $|u\rangle$.

7.3 The Brillouin-Wigner Perturbation Results

From the above considerations, we get the result for the perturbed eigenstates

$$\boxed{|\psi\rangle \approx |u\rangle + \underbrace{\sum_{m\neq u} \frac{\langle m|\hat{W}|u\rangle}{E - E'_m}|m\rangle}_{\phi^{(1)}}\,.} \tag{7.22}$$

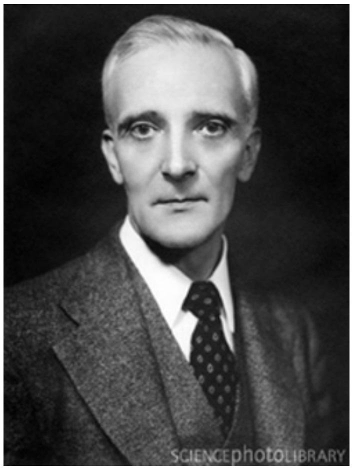


Fig. 7.3: Leon Brillouin, as in the Brillouin function, and the Brillouin zone in solid state physics, is also the 'B' of the WKB approximation. Discovered one of the three fundamental light-matter scattering processes (the other two being Rayleigh and Raman scattering).

The perturbed state vector given by equation 7.25 now is in a way that can be used for calculations. That is because every term on the right side is known, except for the energy E in the denominator. To obtain the perturbed eigenvalues E, we substitute equation 7.2 into the expression for energy $E = E_u + \langle u|W|u\rangle + \langle u|W|\phi\rangle$ to obtain

$$\boxed{E \approx E_u + \underbrace{\langle u|W|u\rangle}_{\Delta E^{(1)}} + \underbrace{\sum_{m\neq u} \frac{|\langle m|\hat{W}|u\rangle|^2}{E - E'_m}}_{\Delta E^{(2)}}\,.} \tag{7.23}$$

This result is called the **Brillouin-Wigner** (BW) perturbation theory. Note that the BW algebraic solution for determining the unknown eigenvalues E require us to solving for it. But for multiple states, the solution would require a high order polynomial, since equation 7.23 is indeed a polynomial. For example, lets say we were looking at a 3-state problem with unperturbed energies E_{u1}, E_{u2}, E_{u3}, and we want to find how eigenvalues of state $u = 2$ got modified by the perturbation. Then, the 2nd energy energy correction has 2 terms, since $m \neq 2$. The equation then becomes a 3rd-order polynomial with three roots, which are the desired eigenvalues.

7.4 Rayleigh-Schrodinger Perturbation Results

The BW perturbation results require us to solve the polynomial equations for the perturbed energies. This can be avoided if we are willing to compromise on the accuracy. If so, the unknown energy term E in the denominator of the 2nd order correction term may be replaced by the unperturbed value, $E \to E_u$. Then the energy eigenvalues are obtained directly from

$$\boxed{E \approx E_u + \langle u|\hat{W}|u\rangle + \sum_{m\neq u} \frac{|\langle m|\hat{W}|u\rangle|^2}{E_u - E'_m},} \tag{7.24}$$

and the eigenfunctions are

$$\boxed{|\psi\rangle \approx |u\rangle + \underbrace{\sum_{m\neq u} \frac{\langle m|\hat{W}|u\rangle}{E_u - E'_m}|m\rangle}_{\phi^{(1)}}\,.} \tag{7.25}$$

This set of perturbed eigenfunction and eigenvalues is called the **Rayleigh-Schrodinger** (RS) perturbation theory. Note that in this form, we know all the terms on the right side. It was first derived by Schrodinger right after his seminal work on the wave equation of electrons. The RS-theory is *not* applicable for understanding perturbation of *degenerate states*, as the denominator $E_n - E'_m$ can go to zero. But BW-theory applies for degenerate states too, and one can always resort back to the degenerate perturbation theory. Schrodinger originally derived this result and referred to Rayleigh's (Figure 7.6) work on classical perturbation theory of the effect of inhomogeneities on the vibration frequencies of mechanical strings. The quantum naming scheme pays homage to the quantum and the classical versions.

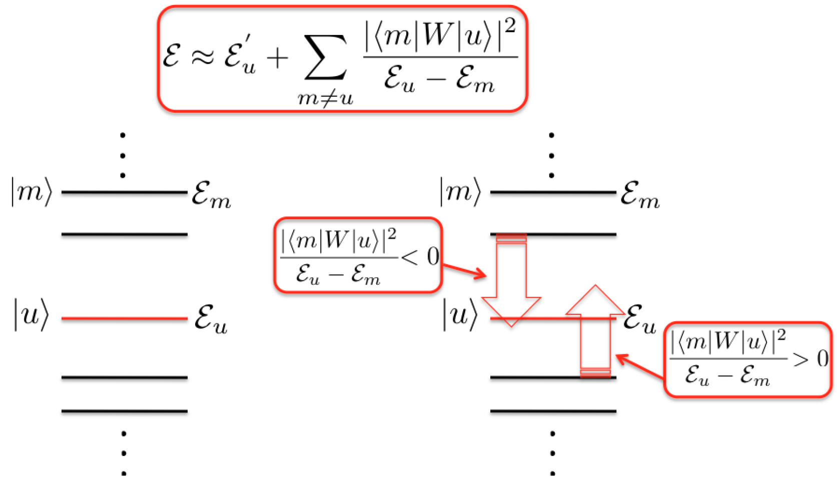

$\mathcal{E} \approx \mathcal{E}_u' + \sum_{m\neq u} \frac{|\langle m|W|u\rangle|^2}{\mathcal{E}_u - \mathcal{E}_m}$
$|m\rangle$
$\mathcal{E}_m$
$|u\rangle$
$\mathcal{E}_u$
$|m\rangle$
$\mathcal{E}_m$
$\frac{|\langle m|W|u\rangle|^2}{\mathcal{E}_u - \mathcal{E}_m} < 0$
$|u\rangle$
$\mathcal{E}_u$
$\frac{|\langle m|W|u\rangle|^2}{\mathcal{E}_u - \mathcal{E}_m} > 0$

Fig. 7.5: Illustration of revel repulsion due to perturbation.

Fig. 7.4: Eugene Wigner, a mathematical physicist who introduced seminal notions of symmetry in atomic spectra, quantum mechanics, and solid-state physics. Wigner was awarded the 1963 Nobel prize in physics. Wigner's memorable statement: 'It is nice that the computer understands the problem. But I would like to understand it too'. Dirac was Wigner's brother-in-law.

In the treatment of degenerate perturbation theory earlier, we discussed the strategy to follow to choose which states to include in the matrix. The last term in the BW or RS perturbation theory results provides the guiding principle. Note that this term goes as the perturbation matrix element *squared*, divided by the energy difference. In the absence of the perturbation, the eigenvectors corresponding to the eigenvalues were orthogonal, meaning they did not 'talk' to each other. The perturbation mixes the states, and makes them talk. The magnitude by which the energy of a state $|u\rangle$ changes due to interactions with *all other states* upon perturbation is $\Delta E^{(2)} \approx \sum_{m\neq u} |W_{mu}|^2/(E_u - E_m')$.

We also note the nature of the interaction. If a state E_u is interacting with states with energies E_m' lower than itself, then $\Delta E^{(2)} > 0$, the perturbation pushes the energy *up*. Similarly, interactions with states with higher energies pushes the energy of state E_u *down*. Thus, the second-order interaction term in perturbation is *repulsive*. Figure 7.5 illustrates this effect schematically. This repulsive interaction is the key to understanding curvatures of energy bands and the relation between effective masses and energy bandgaps of semiconductors. Clearly if two states were non-degenerate and the strength of the perturbation is increased from zero, the energy eigenvalues repel stronger, and the levels go farther apart. Then they cannot cross each other. This is a case of what goes by the name of **no level crossing theorem** in perturbation theory.

In this chapter, we developed the theoretical formalism for handling time-independent perturbation theory. The matrix formalism is well-suited for uncovering the effect of perturbation on eigenvectors and eigenvalues. It works for problems where the unperturbed states are either degenerate, or non-degenerate energy states. For non-degenerate eigenstates, algebraic solutions can be obtained in the Brillouin-Wigner (BW), or the Rayleigh-Schrodinger (RS) theories. The analytical solutions offer further insights into the effect of the perturbation on the physical parameters of interest in the problems. In the next few chapters, we apply both degenerate and non-degenerate perturbation theories to understand electron bandstructure in semiconductors, and its various ramifications.

7.5 The Hellmann Feynman Theorem

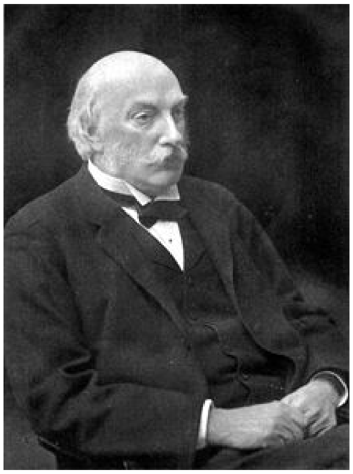


Fig. 7.6: Lord Rayleigh the co-discoverer of Argon, and of Rayleigh scattering of long wavelength light waves from matter that explains why the sky is blue. The perturbation problem is essentially also a scattering problem in disguise, because one can imagine the uperturbed states being scattered into new states because of the perturbing potential. Rayleigh was awarded the 1904 Nobel prize in Physics. J. J. Thomson, the discoverer of the electron, was Rayleigh's student.

The following theorem sometimes is useful for obtaining quick estimates of the magnitude and direction of the shift in energy eigenvalues upon perturbation. Let $\hat{H}^0(\lambda)|k\rangle = E_k(\lambda)|k\rangle$ be the exactly solved problem with normalized eigenstates $|k\rangle$, where the Hamiltonian $\hat{H}^0(\lambda)$ and its resulting eigenvalues $E_k(\lambda)$ depend on a parameter λ. When we change the parameter λ in the Hamiltonian operator $\hat{H}^0(\lambda)$, how do the eigenvalues $E_k(\lambda)$ change?

Because the original eigenstates are orthonormal, we have $\langle k|k\rangle = 1$. Differentiating with respect to the parameter λ, we have $\frac{d}{d\lambda}\langle k|k\rangle = 0$. Now, applying the chain rule for differentiation,

$$\frac{d}{d\lambda}E_k(\lambda) = \frac{d}{d\lambda}\langle k|\hat{H}^0(\lambda)|k\rangle = \langle \frac{d}{d\lambda}k|\hat{H}^0(\lambda)|k\rangle + \langle k|\frac{d}{d\lambda}\hat{H}^0(\lambda)|k\rangle + \langle k|\hat{H}^0(\lambda)|\frac{d}{d\lambda}k\rangle, \tag{7.26}$$

and because $\hat{H}^0(\lambda)|k\rangle = E_k(\lambda)|k\rangle$ and $\langle k|\hat{H}^0(\lambda) = E_k(\lambda)\langle k|$, we get

$$\frac{d}{d\lambda}E_k(\lambda) = \langle k|\frac{d}{d\lambda}\hat{H}^0(\lambda)|k\rangle + E_k(\lambda)\underbrace{[\langle \frac{d}{d\lambda}k|k\rangle + \langle k|\frac{d}{d\lambda}k\rangle]}_{\frac{d}{d\lambda}\langle k|k\rangle=0} \implies \boxed{\frac{dE_k(\lambda)}{d\lambda} = \langle k|\frac{d\hat{H}^0(\lambda)}{d\lambda}|k\rangle}. \tag{7.27}$$

The boxed equation above is the statement of the **Hellmann-Feynman** theorem. It states that we can get the perturbation in the energy eigenvalues due to a parameter λ by finding the inner product of the derivative of the Hamiltonian operator with respect to the variable λ.

7.6 Perturbation Theory Example

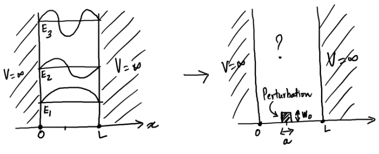


Fig. 7.7: A perturbation to the particle-in-a-box problem.

Consider the particle-in-a-box problem shown in Figure 7.7. Here is the question: we know the exactly solved particle in a box eigenfunctions of state $|n\rangle$: $\langle x|n\rangle = \psi_n(x) = \sqrt{\frac{2}{L}}\sin(n\frac{\pi}{L}x)$ and corresponding eigenvalues $E_n = \frac{\hbar^2}{2m_e}(n\frac{\pi}{L})^2$. If we introduce a tiny perturbation potential of strength $W(x) = +W_0$ over a length $a << L$, and $W_0 << E_1$. Find the perturbed values of the eigenvalues and eigenfunctions for the states $|1\rangle$ and $|2\rangle$.

Brillouin-Wigner and Rayleigh-Schrodinger Perturbation theories: Because the perturbation is[1] $W(x) = W_0[\Theta(x - \frac{L-a}{2}) - \Theta(x - \frac{L+a}{2})]$ the new eigenvalues from the more general BW theory are

$$E = E_u + \langle u|W(x)|u\rangle + \sum_{m\neq u}\frac{|\langle u|W|m|^2}{E - E_u}. \tag{7.28}$$

Denoting $u = n$ as the unperturbed state, we find the first order perturbation shift:

$$W_{nn} = \langle n|W(x)|n\rangle = \frac{2W_0}{L}\int_{\frac{L-a}{2}}^{\frac{L+a}{2}} dx \sin^2(n\frac{\pi}{L}x) = W_0\frac{a}{L}[1 - \cos{(n\pi)}\frac{\sin(n\frac{\pi}{L}a)}{(n\frac{\pi}{L}a)}] \tag{7.29}$$

We note that for even $n = 2, 4, 6, ...$, $\cos(n\pi) = +1$, and for odd $n = 1, 3, 5...$, $\cos(n\pi) = -1$. Because we have assumed the perturbation extends over a small length $a << L$, we can use $\frac{\sin x}{x} \approx 1 - \frac{x^2}{6} + ...$ to get for even n states $\langle n|W(x)|n\rangle \approx W_0\frac{a}{L}\frac{(n\frac{\pi}{L}a)^2}{6} = W_0\frac{n^2\pi^2a^3}{6L^3}$. This is a *very small* change, which goes as $(\frac{a}{L})^3$ compared to the odd states for which we get $\langle n|W(x)|n\rangle \approx W_0\frac{a}{L}[2 + \frac{(n\frac{\pi}{L}a)^2}{6}] \approx 2W_0\frac{a}{L}$, which goes as $\frac{a}{L}$. Why are the odd states

[1] This is a complicated way of writing a rectangular barrier, the meaning is very simple!

so strongly perturbed compared to the even states? The answer lies in the symmetry of the eigenfunctions of the unperturbed states. Because the first order perturbation $\langle n|W(x)|n\rangle$ is due to the perturbation potential $W(x)$ coupling the eigenstates $|n\rangle$ with itself, from Figure 7.8 we see that the perturbation potential is at the maxima of the odd eigenfunctions (odd around $x = 0$, but even around $x = L/2$), but at the nodes or zeroes of the even eigenfunctions (even around $x = 0$, but odd around $x = L/2$). Because the square integral of the even eigenfunctions is non-zero for the small interval $(\frac{L-a}{2} < x < \frac{L+a}{2})$, the even eigenfunctions still get perturbed, but by a minuscule amount. Because the odd eigenfunctions reach their maxima in the same interval, they are perturbed by a large amount.

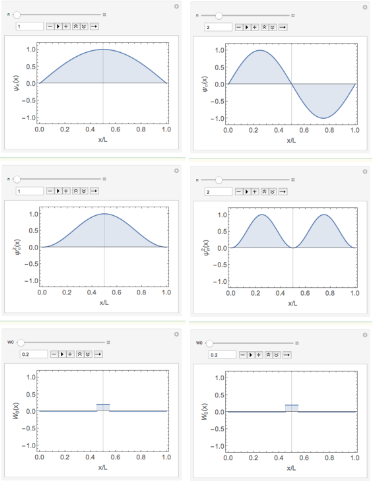


Fig. 7.8: Unperturbed eigenfunctions and probabilities. Symmetry determines if the perturbation affects certain eigenvalues, or not. The $n = 1$ state is strongly affected because its probability peaks at the location of the perturbation. Because the probability of the $n = 2$ state is a minimum at the location of the perturbation, it gets very minimally affected.

The matrix element in the second-order eigenvalue perturbation term $\langle n|W(x)|m\rangle$ is then

$$W_{nm} = \langle n|W(x)|m\rangle = \frac{2W_0}{L}\int_{\frac{L-a}{2}}^{\frac{L+a}{2}} dx \sin(n\frac{\pi}{L}x)\sin(m\frac{\pi}{L}x) \tag{7.30}$$

We can evaluate the integral and get a most general analytical expression which is long and uninspiring. But by invoking symmetry, we can predict the fate of the eigenvalues due to the second order perturbation terms. Consider Figure 7.9. Since the perturbation is symmetric around $x = \frac{L}{2}$, we argue that the integral is zero whenever n is odd and m is even, or if n is even and m is odd. The matrix element is non-zero only when both n, m are odd or even. This means the perturbation potential is capable of only coupling say state $|1\rangle$ with $|3\rangle$, $|5\rangle$, ..., but state $|1\rangle$ cannot couple to (or 'talk to') states $|2\rangle$, $|4\rangle$, ... This is an example of a **selection rule** for coupling.

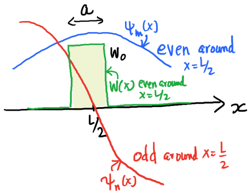


Fig. 7.9: Allowed and forbidden matrix elements obtained from the symmetry of eigenfunctions and the perturbation.

Let us find the matrix elements for the ground state $|1\rangle$. The matrix elements needed are $W_{m1} = \langle m|W(x)|1\rangle = 4W_0\frac{\sin(\frac{m\pi}{2})}{(m^2-1)\pi}[m\cos(\frac{\pi a}{2L})\sin(m\frac{\pi a}{2L}) - \sin(\frac{\pi a}{2L})\cos(m\frac{\pi a}{2L})]$. Because for even m, the term $\sin(\frac{m\pi}{2}) = 0$, state $|1\rangle$ does not couple to any other even state due to the perturbation. For all m =odd states, we make use of the small arguments of the sine and cosine above to estimate $W_{m1} \approx \frac{2}{L}W_0 a$, and write the second order perturbation terms as $\frac{|W_{m1}|^2}{E-E_m}$.

Brillouin-Wigner Perturbation Theory: We then combine the first and second order perturbation terms to write the perturbed eigenvalue to be

$$E \approx E_1 + \frac{2a}{L}W_0 + \sum_{m=3,5,...}\frac{(\frac{2aW_0}{L})^2}{E - [\frac{\hbar^2}{2m_e}(\frac{m\pi}{L})^2 + \frac{2a}{L}W_0]}. \tag{7.31}$$

This is the BW result. To make progress, we must solve the above polynomial equation. Let us denote $E_0 = \frac{2a}{L}W_0$ If we consider only $m = 3$ in the second order sum and neglect $m = 5, 7, ...$, we get $(E - E_1 - E_0)(E - E_3 - E_0) \approx E_0^2$, which give the eigenvalues $E_1' = \frac{1}{2}[(E_1 + E_3) + 2E_0 - \sqrt{(E_3 - E_1)^2 + 4E_0^2}]$, and $E_3' = \frac{1}{2}[(E_1 + E_3) + 2E_0 + \sqrt{(E_3 - E_1)^2 + 4E_0^2}]$. If $E_0 = 0$, we recover the original eigenvalues. Since the unperturbed energy eigenvalues are non-degenerate in this problem, it is illuminating to apply the RS perturbation theory.

Rayleigh-Schrodinger Perturbation Theory: The RS theory from Equation 7.24 gives the perturbed ground state energy

$$E_1' \approx E_1 + E_0 + \sum_{m=3,5,...}\frac{\frac{E_0^2}{E_1}}{1 - m^2 - \frac{E_0}{E_1}} \approx E_1 + E_0 - \frac{E_0^2}{8E_1}. \tag{7.32}$$

Because m^2 increases rapidly with m and far exceeds $1 - \frac{E_0}{E_0}$, we get an acceptable estimate by retaining only the $m = 3$ term: $E_1' \approx E_1 + E_0 - \frac{E_0^2}{8E_1}$. The two perturbation

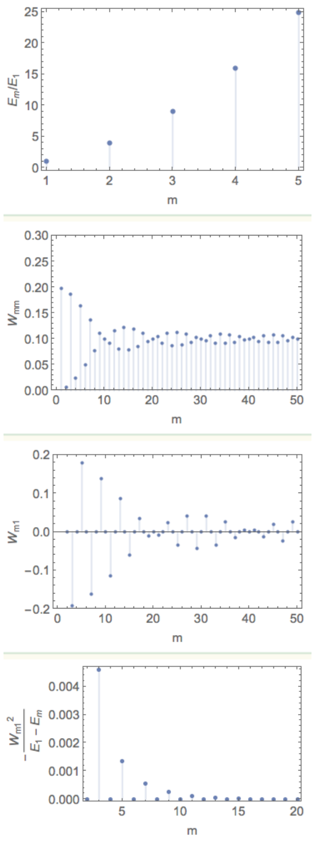


Fig. 7.10: Unperturbed eigenvalues increasing as m^2. Diagonal or 1st-order perturbation W_{mm} alternates between strong and weak before stabilizing for higher states at the spatially averaged perturbation $\approx W_0 \frac{a}{L}$. Off-diagonal matrix elements W_{m1} of the ground state $u = 1$ with higher states m. 2nd order perturbation shift decays rapidly with increasing energy separation.

terms can be now understood: the first-order 'dc' shift due to the perturbation is E_0, and the second order downward energy shift $-E_0^2/8E_1$ is because the perturbation couples state $|1\rangle$ with state $|3\rangle$, the interaction is *always* replusive, and state $|3\rangle$ pushes state $|1\rangle$ energy down. The denominator $8E_1 = E_3 - E_1$ is the energy separation of the two states that are interacting due to the perturbation.

We can also find the change in the eigenfunction using Equation 7.25. To find how the ground state wavefunction has changed, we write

$$|\psi\rangle = |1\rangle + \sum_{m=3,5,...} \frac{E_0}{E_1(1-m^2-\frac{E_0}{E_1})}|m\rangle \implies \psi(x) = \psi_1(x) + \sum_{m=3,5,...} \frac{E_0}{E_1(1-m^2-\frac{E_0}{E_1})}\psi_m(x) \tag{7.33}$$

Because the unperturbed eigenfunctions are $\psi_n(x) = \sqrt{\frac{2}{L}}\sin(n\frac{\pi}{L}x)$, we can make a plot of the new eigenfunction, and specifically its square to find the probability distribution of the ground state. Figure 7.11 shows that the peak at the center $x = L/2$ is reduced from the unperturbed value. The perturbing potential at the center of the well pushes the ground state away from the center.

Matrix Method: As the most general method of perturbation theory, we apply the Matrix method to obtain the perturbed eigenvalues and eigenfunctions. Let us consider the states $m = 1, 2, 3$, and write down the matrix for the perturbed Hamiltonian $H^0 + \hat{W}$:

$$\hat{H}^0 + W = \begin{array}{c|ccc} & |1\rangle & |2\rangle & |3\rangle \\ \hline \langle 1| & E_1 + E_0 & 0 & E_0 \\ \langle 2| & 0 & E_2 & 0 \\ \langle 3| & E_0 & 0 & E_3 + E_0 \end{array}, \tag{7.34}$$

and either by hand, or with a package, find the Eigensystem of this set. We did an example shown in Figure 6.4 in Chapter 6. Now this is a very powerful technique, and matrix algorithms implemented on computers are efficient and fast. Though we are doing this example with a restricted basis of 3 states, it scales to larger basis sets reasonably well.

Figure 7.12 shows the outputs of the matrix perturbation theory with the three-set problem. The perturbed eigenvalues are plotted as a function of the perturbation strength. The values of the strain are greatly exaggerated for this example, but the beauty of the Matrix version is that it is NOT restricted like the BW or RS versions are. The only approximation here is that we are choosing a restricted basis set.

The effect of the perturbation are evident in the movement of the eigenvalues with W_0. State $|2\rangle$ energy is unperturbed, because of symmetry arguments made above. States $|1\rangle$ and $|3\rangle$ are pushed up in energy due to the perturbation potential. However, due to the mutual repulsion between states $|3\rangle$ and $|1\rangle$, state $|3\rangle$ is pushed up, and state $|1\rangle$ is pushed down, leading to noticeable curvatures. Finally, we note that when the perturbation becomes extremely strong, state $|1\rangle$ merges with state $|2\rangle$. This is the limit of a Dirac-delta potential at the center of a particle-in-a-box problem: only the even eigenvalues are allowed because the wavefunction must go to zero at the location of the Dirac delta potential. The shift in the wavefunction and probability distribution for the ground state due to perturbation calculated from the Matrix method is shown in Figure 7.12 too. These are approximate because of the small basis set, but tell the correct story.

Chapter Summary

By combining the **Matrix method**, the **Brillouin-Wigner** method, and the **Rayleigh-Schrodinger** method, we learnt a powerful bag of tricks in this chapter to apply to quantum mechanical problems that are not analytically solvable by dividing the problem to an analytically solvable part, and treating the rest as a perturbation.

Problems

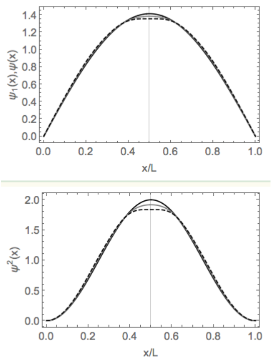


Fig. 7.11: The perturbation pushes some of the probability out from the center as is expected from purely classical grounds.

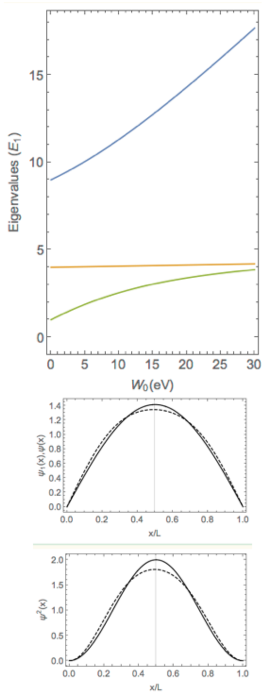


Fig. 7.12: The perturbation of the three lowest eigenvalues neglecting all higher states. The effect on the 2nd state is forbidden by symmetry, and the repulsive interaction between states $m = 1$ and $m = 3$ are evident. The matrix formulation also gives the expansion coefficients for the perturbed wavefunctions from which the new eigenfunctions can be created.

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 8

Electrons in a Crystal get their Bands, Gaps and Masses

Contents

8.1 The free-electron 76
8.2 Periodic perturbation . . 76
8.3 Degenerate Perturbation Theory 77
8.4 Non-degenerate Perturbation Theory 81
8.5 Glimpses of the Bloch Theorem 82
8.6 Non-periodic potentials and scattering 82
Summary 83
Problems 84

Sommerfeld (and Bethe) theory of electrons in a metal could explain most measured properties of a metal. However, even in late 1930s, the theory gave no hint whatsoever about the *existential* question: why are some solids metals, and others insulators? The late 1930s was when most of quantum mechanics was formalized by the likes of Schrodinger, Heisenberg, and Dirac. A rather remarkable yet bewildering logical consequence of the wave nature of electrons when they are put in a periodic potential was discovered by Felix Bloch. Bloch showed that if $V(x+L) = V(x)$, then the wavefunction is of the form $\psi_k(x) = e^{ikx}u_k(x)$, where $u_k(x+L) = u_k(x)$. However, we will come back to Bloch's approach only at the end of this chapter, from a perturbative route.

We discussed in Chapters 3 and 5 how the Pauli exclusion principle led to the elements of the periodic table. To fill the allowed electron shell orbitals allowed by the Schrodinger equation, as we add electrons to the orbitals (and protons and neutrons to the nucleus) to create new elements, every now and then there are situations when a shell becomes completely filled, and the element becomes unreactive. That was the reason behind the existence of the Noble gases He, Ne, Ar, ...

Fig. 8.1: Alan Wilson in 1930s explained how the number of electrons and periodic arrangement of atoms decides if a solid is a metal, a semiconductor, or an insulator. The idea is similar to the formation of open and closed shells for electrons in atoms periodically as the electron number increases. The formation of a crystal causes the formation of bands and gaps for electron energies; partially filled bands are conductive, similar to chemically reactive elements with open shells. If the highest band is completely filled with a significant energy gap, it is an insulator.

If we extend this concept to the *conduction* electron states in a *crystal*, it may be possible that as we increase the electron concentration, the conductivity increases, but at a certain level the conductivity decreases and the 'electron shells' in a crystal 'closes', and the electrons become *inert*, or in other words, we have an insulating solid. This should periodically repeat. Could be the explanation for the periodic occurence of metallic and insulating crystals as we traverse the periodic table? Alan Wilson (Figure 8.1) explained how electrons in solids lead to metallic, semiconducting, or insulating behavior depending on their number, and the presence of the periodic potential due to the atoms in the crystal.

In Chapter 7, we developed the formalism for time-independent perturbation theory. In this chapter, we apply the theory to a free electron perturbed by **a periodic potential**. The results we obtain will highlight most of the fundamental properties of semiconductors. These include their **energy bandstructure** $E(k)$ and opening of **bandgaps** E_g, evolution of **effective masses** $m^\star$ of various bands, work function, interactions between electron states in solids, and the role of **defects** on the interactions between electron states. In brief, the chapter will capture the essence of time-independent semiconductor physics. Much of the following chapters are detailed treatments of increasing levels of sophistication, till we need time-dependent behavior which will require new concepts. The central time-independent phenomena are captured in this chapter. We start from the free electron problem.

8.1 The free-electron

In earlier chapters we have discussed the electron-in-a-ring or free electron in a periodic 1D circle problem in quantum mechanics. The potential term in the Schrodinger equation is zero $V(x) = 0$. The eigenvectors $|k\rangle$ are such that their real-space projection yields the plane wave-function

$$\langle x|k\rangle = \psi_k(x) = \frac{1}{\sqrt{L}}e^{ikx}, \tag{8.1}$$

with corresponding eigenvalues

$$E_0(k) = \frac{\hbar^2 k^2}{2m_0}, \tag{8.2}$$

where $m_0 \sim 9.1 \times 10^{-31}$kg is the free-electron mass. We work in a periodic-boundary condition picture, which requires $\psi(x+L) = \psi(x)$, which requires that the k's are discrete, given by $k_n = (2\pi/L)n$ where n is any integer. We note immediately that a cross-matrix element of the type

$$\langle k_m|k_n\rangle = \int dx \langle k_m|x\rangle\langle x|k_n\rangle = \int_0^L dx \psi^\star_{k_m}(x)\psi_{k_n}(x) = \frac{1}{L}\int_0^L dx e^{i2\pi(n-m)x} = \delta_{n,m} \tag{8.3}$$

is a Kronecker-delta function. This is of course how it should be, since the eigenvectors states $|k_n\rangle$ and $|k_m\rangle$ are mutually orthogonal if (n,m) are different, and the states are normalized to unity. The Hamiltonian matrix is thus diagonal, with the diagonal matrix elements $\langle k|\hat{H}^0|k\rangle = E_0(k)$ given by the free-electron bandstructure in equation 8.2. The off-diagonal elements $\langle k_m|\hat{H}^0|k_n\rangle = E_n\langle k_m|k_n\rangle$ are zero because of equation 8.3.

8.2 Periodic perturbation

In a crystalline solid, the electron experiences a periodic potential. To model the situation for the electron on the ring, let's add a perturbation to the free electron in the form of a periodic potential. The perturbation potential is

$$W(x) = -2U_G\cos(Gx) = -U_G(e^{iGx} + e^{-iGx}), \tag{8.4}$$

where U_G is the 'strength' in units of energy, and $G = 2\pi/a$, where a is the lattice-constant of the perturbation. This periodic potential is shown in Fig 8.2. The lowest energy of a *classical* particle in this potential landscape is clearly $-2U_G$, at the bottom of a valley. The new Hamiltonian is then $\hat{H} = \hat{H}^0 + W(x) = -\frac{\hbar^2}{2m_0}\frac{\partial^2}{\partial x^2} - 2U_G\cos(Gx)$. In principle this 1D Schrodinger-equation can be solved numerically to a large degree of accuracy *directly* without perturbation theory. But we are going to apply perturbation theory to highlight the insights it affords.

We can find the entire Hamiltonian matrix if we find the matrix elements $\langle k_2|\hat{H}^0 + W(x)|k_1\rangle = E_0(k_1)\delta_{k_1,k_2} + \langle k_2|W(x)|k_1\rangle$. The first term is the unperturbed diagonal matrix element, and the second term is due to the perturbation. The perturbation matrix element evaluates to

$$\langle k_2|W(x)|k_1\rangle = -\frac{U_G}{L}\int_0^L dx e^{i(k_1-k_2)x}(e^{iGx} + e^{-iGx}) = -U_G\delta_{k_1-k_2,\pm G}. \tag{8.5}$$

The Kronecker-delta implies that the perturbation only couples states $|k_1\rangle$ and $|k_2\rangle$ if their wavevector difference is $k_1 - k_2 = \pm G$, the reciprocal lattice vector of the perturbing

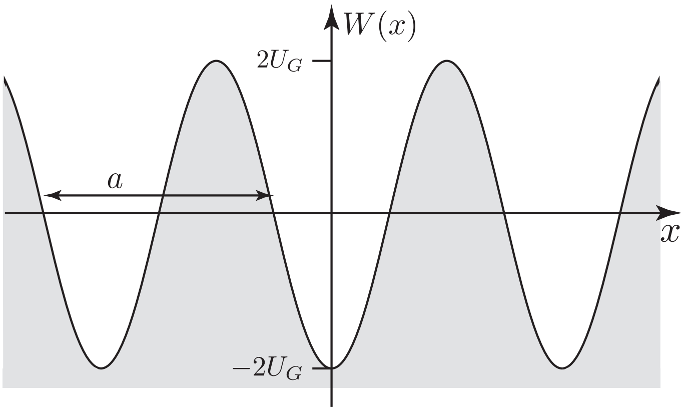


Fig. 8.2: A periodic potential $W(x) = -2U_G \cos(Gx)$ acts as a perturbation to the free electron.

potential. Recall from chapter 7 that we can find the perturbed eigenvalues by the matrix method, which works *both* for degenerate and non-degenerate states. But if we were to consider all the $|k\rangle$ states, the matrix would be ∞-dimensional. So we should choose a restricted set for identifying the eigenvalues.

8.3 Degenerate Perturbation Theory

It is clear from equation 8.5 that a state $|k\rangle$ will interact due to the periodic perturbation with only two other states $|k+G\rangle$ and $|k-G\rangle$ directly. This will require us to solve a 3×3 Hamiltonian for the three states $|k-G\rangle$, $|k\rangle$, and $|k+G\rangle$. But also recall in chapter 7 the result of non-degenerate perturbation theory told us that the changes in eigenvalues for states widely separated in energy goes as $|W_{12}|^2/(E_1 - E_2)$. So the states that interact *most strongly* due to the perturbation must be close (or degenerate) in energy, but their wavevectors should still follow $k_1 - k_2 = \pm G$. Clearly, two such states are $|+G/2\rangle$ and $|-G/2\rangle$. This is illustrated in Fig 8.3. To locate states that have non-zero matrix elements, one has to imagine sliding the double-headed arrow of length G along the $k-$axis. Two situations are shown, one when the unperturbed states are degenerate, and one when they are not. Also remember the repulsive nature of the interaction: in Fig 8.3 we expect state $|k_1\rangle$ to be pushed *down*, and state $|k_2\rangle$ to be pushed *up* due to their mutual interaction.

The unperturbed eigenvalue of the two degenerate states is $E_0(G/2) = \hbar^2 G^2/8m_0 = F$. Clearly this is a case for the application of *degenerate* perturbation theory[1]. The problem is rather simple, since the Hamiltonian is a 2×2 matrix:

$$\hat{H}^0 + W = \begin{array}{c c} & \begin{array}{cc} |+\frac{G}{2}\rangle & |-\frac{G}{2}\rangle \end{array} \\ \begin{array}{c} \langle +\frac{G}{2}| \\ \langle -\frac{G}{2}| \end{array} & \left(\begin{array}{cc} F & -U_G \\ -U_G & F \end{array} \right) \end{array}, \tag{8.6}$$

where we write out the ket and bra states explicitly to highlight where the matrix elements come from. The eigenvalues of this matrix are obtained by solving the determinant of the matrix: $(F-E)^2 - U_G^2 = 0$, which yields $E_\pm = F \pm U_G$. This implies the degenerate unperturbed states $E_0(+G/2) = E_0(-G/2) = F$ have now been split to two energies E_+ and E_- with the difference $E_+ - E_- = 2U_G$ by the periodic perturbation. This is the opening of a *bandgap* in the allowed energies for the electron, and is highlighted in Fig 8.3.

[1] We will see later that the Brillouin-Wigner (BW) non-degenerate perturbation theory also can give

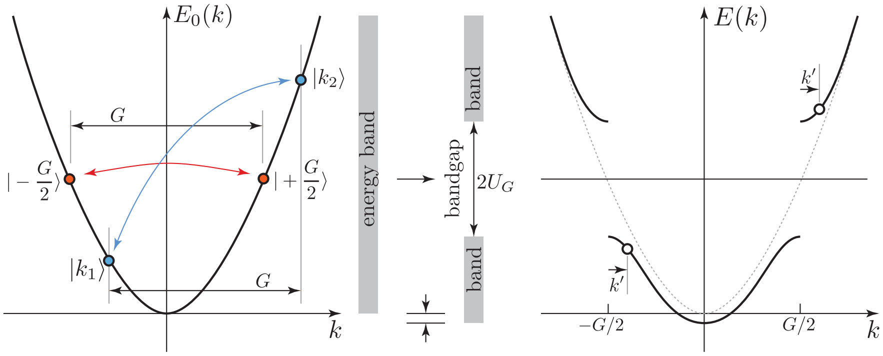


Fig. 8.3: Bandgap opening in the energy spectrum of a free electron upon perturbation by a periodic potential.

We note here that the general eigenvalues of the 2×2 Hamiltonian matrix

$$\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \tag{8.7}$$

are

$$E_{\pm} = \frac{H_{11}+H_{22}}{2} \pm \sqrt{(\frac{H_{11}-H_{22}}{2})^2 + |H_{12}|^2}, \tag{8.8}$$

the corresponding eigenvectors are

$$\left[\begin{array}{c} a_1 \\ a_2 \end{array}\right]_{\pm} = \left[\begin{array}{c} \frac{H_{12}}{\sqrt{|H_{12}|^2+(E_{\pm}-H_{11})^2}} \\ \frac{E_{\pm}-H_{11}}{\sqrt{|H_{12}|^2+(E_{\pm}-H_{11})^2}} \end{array}\right] \tag{8.9}$$

Since we expand the perturbed states as $|\psi\rangle = \sum_n a_n |n\rangle$, for the degenerate states, we obtain the perturbed eigenvectors as

$$|\pm\rangle = a_{1\pm}|+\frac{G}{2}\rangle + a_{2\pm}|-\frac{G}{2}\rangle. \tag{8.10}$$

For the degenerate states we get $a_{1+} = -1/\sqrt{2}$ and $a_{2+} = +1/\sqrt{2}$, and $a_{1-} = -1/\sqrt{2}$ and $a_{2-} = -1/\sqrt{2}$. The identification of the coefficients helps us convert the perturbed eigenvectors into the eigenfunctions

$$\langle x|+\rangle = \psi_+(x) = (-\frac{1}{\sqrt{2}})\cdot(\frac{e^{i\frac{G}{2}x}}{\sqrt{L}}) + (+\frac{1}{\sqrt{2}})\cdot(\frac{e^{-i\frac{G}{2}x}}{\sqrt{L}}) = -i\sqrt{\frac{2}{L}}\sin(\frac{G}{2}x), \tag{8.11}$$

and

$$\langle x|-\rangle = \psi_-(x) = (-\frac{1}{\sqrt{2}})\cdot(\frac{e^{i\frac{G}{2}x}}{\sqrt{L}}) + (-\frac{1}{\sqrt{2}})\cdot(\frac{e^{-i\frac{G}{2}x}}{\sqrt{L}}) = -\sqrt{\frac{2}{L}}\cos(\frac{G}{2}x). \tag{8.12}$$

This is illustrated in Fig 8.4. Note now the properties of $|\psi_+(x)|^2 = (2/L)\sin^2(Gx/2)$ and $|\psi_-(x)|^2 = (2/L)\cos^2(Gx/2)$. The probability densities for the higher energy states

the same result.

$E_+ = F + U_G$ go as $\sin^2(Gx/2)$, meaning they peak at the highest points of the perturbing potential. The high potential energy is responsible for the high net energy of these states. Similarly, the lower energy states $E_- = F - U_G$ pile up in the valleys, and consequently have lower energies. Note that due to the perturbation, the new eigenfunctions of the degenerate states *no longer* have a uniform probability distribution in space.

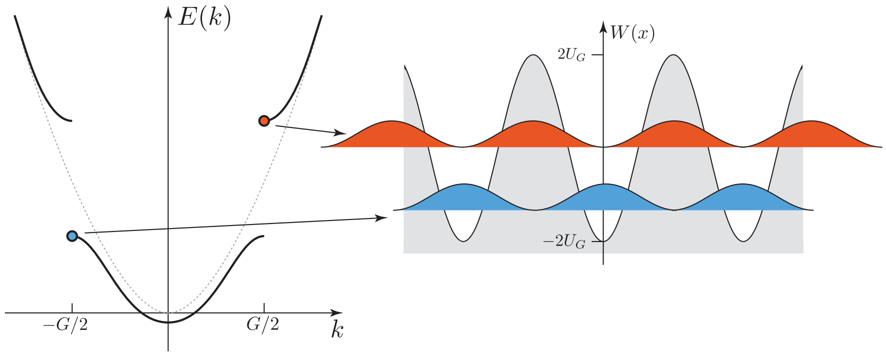


Fig. 8.4: Probability pileups of band-edge states.

But what about states that are not degenerate? Let's look at the states $|k_2\rangle = |\frac{G}{2} + k'\rangle$ and $|k_1\rangle = |-\frac{G}{2} + k'\rangle$, for example those shown in Fig 8.3. By tuning the magnitude of k', we can move as close to the $\pm G/2$ states as possible. The perturbed Hamiltonian is

$$\hat{H}^0 + W = \begin{array}{c|cc} & |+\frac{G}{2}+k'\rangle & |-\frac{G}{2}+k'\rangle \\ \hline \langle +\frac{G}{2}+k'| & E_0(+\frac{G}{2}+k') & -U_G \\ \langle -\frac{G}{2}+k'| & -U_G & E_0(-\frac{G}{2}+k') \end{array}, \tag{8.13}$$

where we write the diagonal unperturbed eigenvalues as

$$E_0(\pm\frac{G}{2} + k') = \underbrace{\frac{\hbar^2 G^2}{8m_0}}_{F} + \underbrace{\frac{\hbar^2 k'^2}{2m_0}}_{\mathcal{E}(k')} \pm \underbrace{\frac{\hbar^2 G k'}{2m_0}}_{2\sqrt{F\mathcal{E}(k')}} = F + \mathcal{E}(k') \pm 2\sqrt{F\mathcal{E}(k')}. \tag{8.14}$$

The eigenvalues then are obtained from equation 8.8 as

$$E_\pm(k') = F + \mathcal{E}(k') \pm \sqrt{4F\mathcal{E}(k') + U_G^2} \approx F + \mathcal{E}(k') \pm U_G(1 + \frac{2F\mathcal{E}(k')}{U_G^2}), \tag{8.15}$$

where we have expanded the square root term using $(1+x)^n \approx 1 + nx + \ldots$ for $x << 1$ assuming $4F\mathcal{E}(k')/U_G^2 << 1$. The energy dispersion then becomes

$$E_\pm(k') \approx (F \pm U_G) + (1 \pm \frac{2F}{U_G})\frac{\hbar^2 k'^2}{2m_0}, \tag{8.16}$$

from where we choose the + sign as a 'conduction' band with lowest energy $E_c(0) = F + U_G$, and the − sign as the 'valence' band with highest energy $E_v(0) = F - U_G$. We rewrite the energy dispersions as

$$\begin{aligned} E_c(k') &\approx E_c(0) + \tfrac{\hbar^2 k'^2}{2m_c^\star} \\ E_v(k') &\approx E_v(0) + \tfrac{\hbar^2 k'^2}{2m_v^\star} \end{aligned} \tag{8.17}$$

where the conduction band *effective* mass is $m_c^\star = \frac{m_0}{1+\frac{2F}{U_G}}$, and the valence band *effective* mass is $m_v^\star = \frac{m_0}{1-\frac{2F}{U_G}}$. We note immediately that the *effective* mass of carriers at the band-edges is different from the mass of the free-electron. The conduction band edge effective mass is *lower* than the free electron mass; the electron moves as if it is lighter. If we assume that $U_G << F$, we can neglect the 1 in the denominator, and we get the interesting result that $m_c^\star \sim (U_G/2F)m_0$, that is, **the effective mass is proportional to the energy bandgap**. We will see later in later chapters in the $\mathbf{k}\cdot\mathbf{p}$ theory that for most semiconductors, this is an excellent rule of thumb for the conduction band effective mass.

The valence band effective mass under the same approximation is $m_v^\star \sim -(U_G/2F)m_0$, i.e., it is *negative*. This should not bother us at least mathematically, since it is clear that the bandstructure curves *downwards* in Fig 8.3, so its curvature is negative. Physically, it means that the electron in the valence band moves in the opposite direction to an electron in the conduction band in the same $|k\rangle$ state. This is clear from the *group velocity* $v_g = \hbar^{-1}dE(k)/dk$: the slopes of the states are opposite in sign.

Are there $k-$states other than $|\pm G/2\rangle$ at which energy gaps develop due to the perturbation? Let's examine the states $|\pm G\rangle$, with unperturbed energy $4F$. Clearly, $k_2 - k_1 = 2G$, so there is no *direct* interaction between the states. But an *indirect* interaction of the form $|-G\rangle \leftrightarrow |0\rangle \leftrightarrow |+G\rangle$ is possible. This is illustrated in Fig 8.5. The eigenvalues for such an interaction are found by diagonalizing the 3×3 perturbation Hamiltonian

$$\hat{H}^0 + W = \begin{array}{c} \\ \langle -G| \\ \langle 0| \\ \langle +G| \end{array} \begin{array}{c} \begin{array}{ccc} |-G\rangle & 0 & |+G\rangle \end{array} \\ \left(\begin{array}{ccc} 4F & -U_G & 0 \\ -U_G & 0 & -U_G \\ 0 & -U_G & 4F \end{array}\right) \end{array}, \tag{8.18}$$

which yields the perturbed eigenvalues $4F, 2F \pm \sqrt{4F^2 + 2U_G^2}$. If the perturbation potential is weak, i. e., $U_G << F$, then we can expand the square root to get the three eigenvalues $4F, 4F + U_G^2/2F, -U_G^2/2F$. We note that there indeed is a splitting of the $|\pm G\rangle$ states, with an energy bandgap $U_G^2/2F$. Similarly, gaps will appear at $\pm mG/2$, due to indirect interactions $|-mG/2\rangle \leftrightarrow |-(m/2+1)G\rangle \ldots \leftrightarrow |+mG/2\rangle$, with a bandgap that scales as U_G^m. For example, the indirect interaction $|-3G/2\rangle \leftrightarrow |-G/2\rangle \leftrightarrow |+G/2\rangle \leftrightarrow |+3G/2\rangle$ is depicted schematically in Fig 8.5.

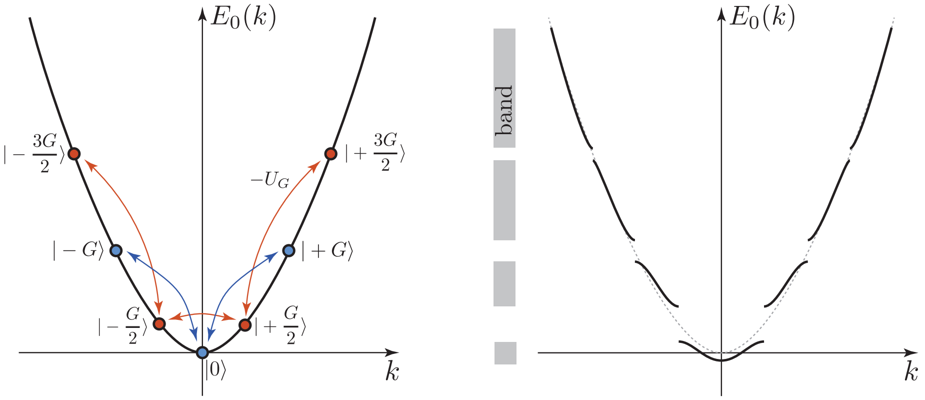


Fig. 8.5: Indirect coupling via intermediate states. Each coupling has a strength $-U_G$.

We also note that the intermediate state $|k = 0\rangle$ which had a zero unperturbed energy now has been pushed down, and has a negative energy of $-U_G^2/2F$. Thus the ground state energy of the electron is now *negative*, implying it is energetically favorable that the electron be in this state. This idea develops into the concept of a work-function of a solid: it takes energy to kick out an electron from the ground state into the free-electron state, which has a minimum of zero. The work-function is schematically illustrated in Fig 8.3.

8.4 Non-degenerate Perturbation Theory

A number of results that we obtained in the previous section may be obtained using non-degenerate perturbation theory. Recall non-degenerate perturbation theory in chapter 7 Equation 7.23, provided the Brillouin-Wigner (BW) result

$$E \approx E_u + \langle u|W|u\rangle + \sum_{m\neq u} \frac{|\langle m|\hat{W}|u\rangle|^2}{E - E'_m}, \tag{8.19}$$

with the Rayleigh-Schrodinger result obtained by simply replacing $E \to E_u$ on the right side. Let us investigate whether we can apply it to the electron in a periodic potential problem.

If we apply the BW-theory to the states $|u\rangle = |\pm G/2\rangle$, we identify $E_u = F$, $\langle u|\hat{W}|u\rangle = 0$, and $\langle -G/2|\hat{W}| + G/2\rangle = -U_G$, the sum in the RHS of equation 8.19 has just one term, and we get

$$E \approx F + \frac{U_G^2}{E - F} \implies E \approx F \pm U_G, \tag{8.20}$$

which actually yields the same result as obtained by degenerate perturbation theory in the last section. The two degenerate states are split, with a gap of $2U_G$. This is an advantage of the BW-theory: it works even for degenerate states, though it is typically classified under non-degenerate perturbation theory. Note that we had to solve the same quadratic equation as the 2×2 matrix in the degenerate theory. They are the same thing. The disadvantage of the BW theory is that it requires us to solve for the roots of a polynomial equation.

Clearly the RS-theory

$$E \approx E_u + \langle u|W|u\rangle + \sum_{m\neq u} \frac{|\langle m|\hat{W}|u\rangle|^2}{E_u - E'_m}, \tag{8.21}$$

cannot be applied to degenerate states, since the the denominator in the sum on the RHS will become zero. But it is well-suited for non-degenerate states. For example, if we ask the question how is state $|0\rangle$ perturbed by its interaction with states $|-G\rangle$ and $|+G\rangle$, we get

$$E \approx 0 + 0 + \frac{U_G^2}{0 - 4F} + \frac{U_G^2}{0 - 4F} = -\frac{U_G^2}{2F}, \tag{8.22}$$

which is the approximate result we had obtained by diagonalizing the perturbation matrix in equation 8.18. For small perturbations, this result is a good approximation. But if U_G increases, it is easy to see that the minimum energy $-U_G^2/2F$ can become lower than the classically minimum energy allowed in the system, which is $-2U_G$. This should be clear from Fig 8.2. The minimum energy allowed for the electron should be *larger* than $-2U_G$ because of quantum confinement, implying $U_G << 4F$.

Application of the BW theory removes this restriction, since it requires the solution of

$$E \approx 0 + 0 + \frac{U_G^2}{E - 4F} + \frac{U_G^2}{E - 4F} = -\frac{2U_G^2}{E - 4F}, \tag{8.23}$$

which yields the same result as the non-degenerate matrix method for state $|0\rangle$: $E \approx 2F - \sqrt{4F^2 + 2U_G^2}$. This root is clearly always greater than $-2U_G$, since it asymptotically approaches $-\sqrt{2}U_G$ when U_G is large. But note that too large a U_G compared to F makes the 'perturbative' treatment not valid in the first place.

8.5 Glimpses of the Bloch Theorem

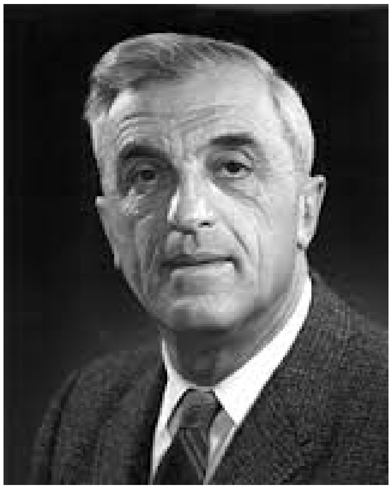


Fig. 8.6: Felix Bloch showed mathematically that electron waves can propagate in a crystal with no scattering, by introducing a wavefunction that electrons experiencing a periodic potential must satisfy. Bloch was awarded the Nobel Prize in physics in 1952 for his work on nuclear magnetic resonance.

Consider the free electron state $|k\rangle$ with real-space projection $\psi_k(x) = \langle x|k\rangle = e^{ikx}/\sqrt{L}$. Due to the periodic potential $W(x) = -2U_G \cos(Gx)$, this state couples to the states $|k+G\rangle$ and $|k-G\rangle$. What is the perturbed *wavefunction* in real space?

From Equation 7.25 in chapter 7, we write the perturbed state vector $|k'\rangle$ as

$$|k'\rangle \approx |k\rangle + \frac{\langle k+G|W|k\rangle}{E(k) - E(k+G)}|k+G\rangle + \frac{\langle k-G|W|k\rangle}{E(k) - E(k-G)}|k-G\rangle \tag{8.24}$$

where we have used the Rayleigh-Schrodinger version. The matrix elements are $-U_G$; projecting the perturbed state on $\langle x|$ we get the perturbed wavefunction to be

$$\psi_{k'}(x) = \langle x|k'\rangle \approx \frac{e^{ikx}}{\sqrt{L}} - \frac{U_G}{E(k) - E(k+G)}\frac{e^{i(k+G)x}}{\sqrt{L}} - \frac{U_G}{E(k) - E(k-G)}\frac{e^{i(k-G)x}}{\sqrt{L}} \tag{8.25}$$

from where we split off e^{ikx} to write the wavefunction as

$$\psi_{k'}(x) \approx e^{ikx} \cdot \underbrace{\left[\frac{1}{\sqrt{L}} - \left(\frac{U_G}{E(k) - E(k+G)}\right)\frac{e^{iGx}}{\sqrt{L}} - \left(\frac{U_G}{E(k) - E(k-G)}\right)\frac{e^{-iGx}}{\sqrt{L}}\right]}_{u_k(x)}. \tag{8.26}$$

Note that the wavefunction is of the form $e^{ikx}u_k(x)$, where the function $u_k(x)$ has the property $u_k(x+a) = u_k(x)$, because $e^{\pm iGa} = 1$. This is really the statement of the Bloch theorem: the eigenfunctions for an electron in the presence of a periodic potential can be written in the form $\psi_k(x) = e^{ikx}u_k(x)$, where $u_k(x+a) = u_k(x)$ has the same periodicity as the potential. A more complicated periodic potential such as $W(x) = -2[U_{G_1}\cos(G_1 x) + U_{G_2}\cos(G_2 x) + ...]$ will lead to more couplings, and create more terms in $u_k(x)$, but the Bloch decomposition of the wavefunction in Equation 8.26 will still remain true. We call this a 'glimpse' of the Bloch theorem because of the '$\approx$' sign in Equation 8.26; in the next chapter this sign will be rigorously turned into an equality. The Bloch theorem is a *non-perturbative* result: it does not depend on the strength of the periodic potential. But of course we just saw it naturally *emerge* as a result from perturbation theory.

8.6 Non-periodic potentials and scattering

We make a few observations of the material covered in this chapter. First, the application of a periodic potential $-2U_G\cos(Gx)$ of reciprocal lattice vector G could only directly couple states that followed $k_2 - k_1 = \pm G$. This caused the appearance of bandgaps due to *direct* interaction at states $|k\rangle = |\pm G/2\rangle$. But due to *indirect* interactions, bandgaps also appeared at $|\pm mG/2\rangle$. If the periodic potential instead was $W(x) = -2[U_{G_1}\cos(G_1 x) + U_{G_2}\cos(G_2 x)]$, we expect direct gaps at more $k-$points, and more direct and indirect coupling of states. The nature of the periodic potential will thus determine the bandstructure. We show this schematically in Figure 8.7

If instead of a periodic potential, we had a localized potential, say $W(x) = V_0 e^{-x/x_0}$, then we can Fourier-expand the potential to obtain $W(x) = \sum_G U_G \cos(Gx)$, and the

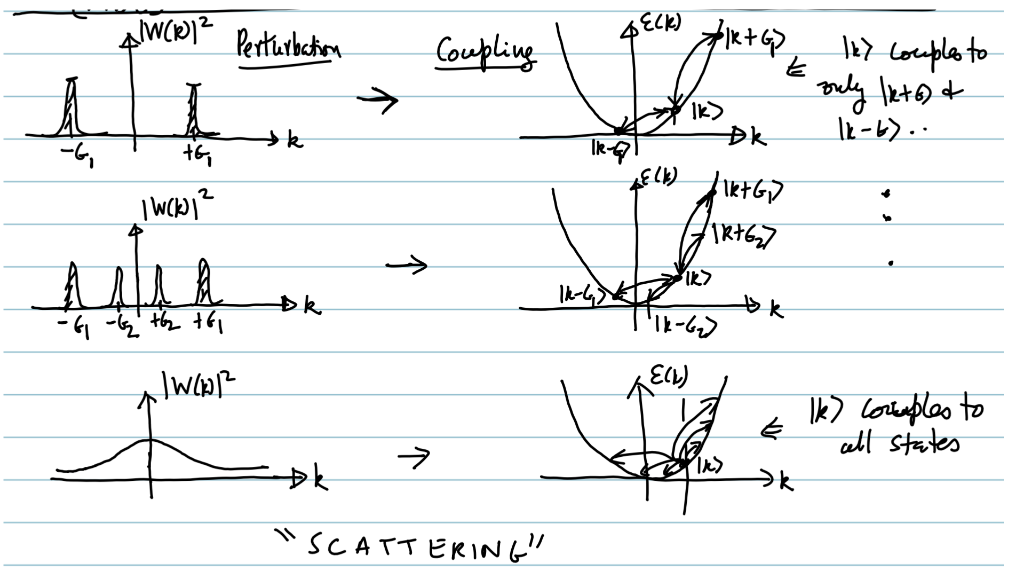


Fig. 8.7: Periodic potentials only scatter states separated by specific G values, and thus open bandgaps at specific k values because they have spectral weight only for specific k's. Non-periodic potentials on the other hand can scatter a state $|k\rangle$ into several states depending on the weight of the potential in the $k-$space.

expansion coefficients will dictate the strength of the couplings. We immediately note that since a localized potential will require a large number of G's, it will effectively couple a wide range of $k-$states, as shown in Figure 8.7. This is why any *deviation* from periodicity will couple a continuum of $k-$states, a phenomena that is responsible for scattering and localization.

Applications of non-degenerate and degenerate perturbation theory can explain a host of phenomena in semiconductors, and other quantum systems. In this chapter, we applied the techniques to the 'toy-model' of an electron in a 1D periodic potential. In the next chapter, we investigate this technique to develop a rather useful model for the electronic bandstructure of realistic semiconductors.

Chapter Summary

- A free electron can occupy states that have a continuum of energy eigenvalues.
- In a periodic potential, free electron states whose wavelengths are resonant with the periodic potential undergo diffraction and scattering, and form standing waves.
- The strong interaction opens bandgaps by forbidding propagating states of electrons within a band of energies, opening bandgaps.
- The total number of states do not change due to the periodic potential. Thus, the states bunch up, and the bands curve near the edges of the gap.
- The curvature of the bands near the gap are called the effective masses.
- Because of level repulsion, larger the gap, larger the effective mass.
- Because periodic potentials couple state $|k\rangle$ with several other states $|k \pm G\rangle$ where G is the reciprocal lattice vector, the new allowed states of the electron take the form of a Bloch function.

- Periodic potentials scatter specific electron states because of the large spectral potential at specific $k = G$.
- Aperiodic, localized, or random potentials scatter electrons into several states.

Problems

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 9

The k dot p Bandstructure

Contents

9.1 Perturbation theory . . . 85
9.2 Symmetry 86
9.3 $\mathbf{k}\cdot\mathbf{p}$ theory 88
9.4 No spin-orbit interaction . 88
9.5 With spin-orbit interaction 91
9.6 Further reading 93

9.1 Perturbation theory

For the $\mathbf{k}\cdot\mathbf{p}$ theory of semiconductor bandstructure, we recap the results of time-independent perturbation theory. There are two levels of the solution - the first is degenerate perturbation theory, and the next non-degenerate theory. For $\mathbf{k}\cdot\mathbf{p}$ theory, the non-degenerate case is important.

Assume that we have solved the Schrodinger equation for a particular potential with Hamiltonian $H^{(0)}$

$$H^{(0)}|n\rangle = \varepsilon_n^{(0)}|n\rangle, \tag{9.1}$$

and obtained the eigenfunctions $|n\rangle$ and eigenvalues $\varepsilon_n^{(0)}$. Now let us see how the solution set (eigenfunction, eigenvalue) changes for a potential that differs from the one we have solved for by a small amount. Denote the new Hamiltonian by $H = H^{(0)} + W$, where W is the perturbation.

If the eigenvalues are non-degenerate, the *first order* energy correction is given by

$$\Delta\varepsilon_n^{(1)} \approx \langle n|W|n\rangle, \tag{9.2}$$

and there is no correction (to first order, in the absence of non-diagonal matrix elements) in the eigenfunction. This is just the diagonal matrix element of the perturbing potential. The second order correction arises from from non-diagonal terms; the energy correction is given by

$$\Delta\varepsilon_n^{(2)} \approx \sum_{m\neq n} \frac{|\langle n|W|m\rangle|^2}{\varepsilon_n^{(0)} - \varepsilon_m^{(0)}}, \tag{9.3}$$

where $|m\rangle$ are all other eigenfunctions. The correction to the eigenfunction is

$$|p\rangle = |n\rangle + \sum_{m\neq n} \frac{\langle m|W|n\rangle}{\varepsilon_n - \varepsilon_m}|m\rangle. \tag{9.4}$$

Thus, the total perturbed energy is given by

$$\varepsilon_n \approx \varepsilon_n^{(0)} + \Delta\varepsilon_n^{(1)} + \Delta\varepsilon_n^{(2)} = \varepsilon_n^{(0)} + \langle n|W|n\rangle + \sum_{m\neq n} \frac{|\langle n|W|m\rangle|^2}{\varepsilon_n^{(0)} - \varepsilon_m^{(0)}}, \tag{9.5}$$

and the perturbed eigenfunction is given by the the equation before last.

Some more facts will have a direct impact on bandstructure calculation by $\mathbf{k}\cdot\mathbf{p}$ method. The total second-order perturbation $\Delta\varepsilon_n^{(2)}$ arises due to the *interaction* between different eigenvalues. Whether interaction between states occurs or not is determined by the matrix elements $\langle n|W|m\rangle$; if it vanishes, there is no interaction. Whether the states vanish or not can typically be quickly inferred by invoking the *symmetry* properties of the eigenfunctions and the perturbing potential W.

Let us look at the effect of interaction of a state with energy $\varepsilon_n^{(0)}$ with all other eignestates. Interacting states with energies $\varepsilon_m^{(0)}$ higher than $\varepsilon_n^{(0)}$ will lower the energy ε_n by contributing a negative term; i.e., they push the energy down. Similarly, states with energies lower than $\varepsilon_n^{(0)}$ will push it up. The magnitude of interaction scales inversely with the difference in energies; so, the strongest interaction is with the nearest energy state.

This is all the basic results that we need for $\mathbf{k}\cdot\mathbf{p}$ theory. The last homework that needs to be done is a familiarity with the consequences of symmetry, which we briefly cover now.

9.2 Symmetry

A brief look at the symmetry properties of the eigenfunctions would greatly simplify solving the final problem, and greatly enhance our understanding of the evolution of bandstructure. First, we start by looking at the energy eigenvalues of the individual *atoms* that constitute the semiconductor crystal. All semiconductors have tetrahedral bonds that have sp^3 hybridization. However, the individual atoms have the outermost (valence) electrons in in s- and p-type orbitals. The symmetry (or geometric) properties of these orbitals are made most clear by looking at their angular parts -

$$s = 1 \tag{9.6}$$

$$p_x = \frac{x}{r} = \sqrt{3}\sin\theta\cos\phi \tag{9.7}$$

$$p_y = \frac{y}{r} = \sqrt{3}\sin\theta\sin\phi \tag{9.8}$$

$$p_z = \frac{z}{r} = \sqrt{3}\cos\theta. \tag{9.9}$$

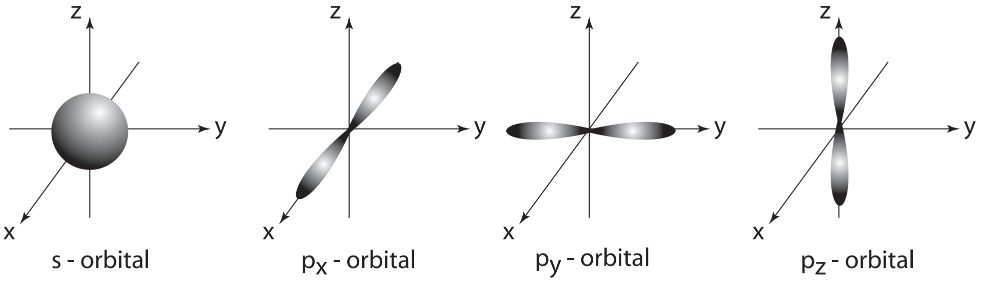


Fig. 9.1: s- and p-orbitals of atomic systems. The s-orbital is spherical, and hence symmetric along all axes; the p-orbitals are antisymmetric or odd along the direction they are oriented - i.e., the p_x orbital has two lobes - one positive, and the other negative.

The spherical s-state and the p-type lobes are depicted in Figure 2. Let us denote the states by $|S\rangle, |X\rangle, |Y\rangle, |Z\rangle$.

Once we put the atoms in a crystal, the valence electrons hybridize into sp^3 orbitals that lead to tetrahedral bonding. The crystal develops its own bandstructure with gaps and allowed bands. For semiconductors, one is typically worried about the bandstructure of the conduction and valence bands only. It turns out that the states near the band-edges

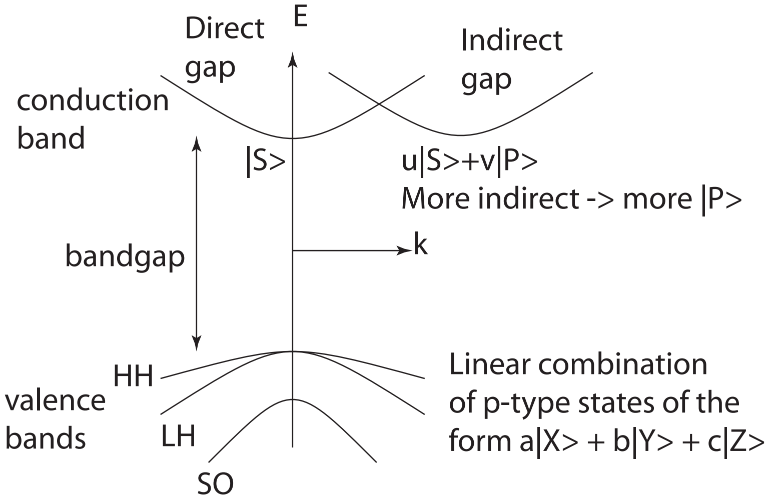


Fig. 9.2: The typical bandstructure of semiconductors. For direct-gap semiconductors, the conduction band state at $\mathbf{k} = 0$ is s-like. The valence band states are linear combinations of p-like orbitals. For indirect-gap semiconductors on the other hand, even the conduction band minima states have some amount of p-like nature mixed into the s-like state.

behave very much like the the $|S\rangle$ and the three p-type states that they had when they were individual atoms.

For direct-gap semiconductors, for states near the conduction-band minimum ($\mathbf{k} = 0$), the Bloch lattice-function $u_c(\mathbf{k}, \mathbf{r}) = u_c(0, \mathbf{r})$ possesses the same symmetry properties as a $|S\rangle$ state[1]. In other words, it has spherical symmetry. The states at the valence band maxima for all bands, on the other hand, have the symmetry of p-orbitals. In general, the valence band states may be written as *linear combinations* of p-like orbitals. Figure 3 denotes these properties. So, we see that the Bloch lattice-functions retain much of the symmetries that the atomic orbitals possess. To put it in more mathematical form, let us say that we have the following Bloch lattice-functions that possess the symmetry of the s- and p_x, p_y, p_z-type states - $u_s, u_x, u_y, \& u_z$. Then, we make the direct connection that u_c is the same as u_s, whereas the Bloch lattice-functions of the valence bands u_v^s are linear combinations of $u_x, u_y, \& u_z$.

Without even knowing the exact nature of the Bloch lattice-functions, we can immediately say that the matrix element between the conduction band state and *any* valence band state is

$$\langle u_c | u_v \rangle = 0, \tag{9.10}$$

i.e., it vanishes. This is easily seen by looking at the orbitals in Figure 2; the p-states are odd along one axis and even along two others; however, the s-states are even. So, the product, integrated over all unit cell is zero. Note that it does not matter which valence band we are talking about, since all of them are linear combinations of p-orbitals.

Next, we look at the momentum-matrix element, $\langle u_c | \mathbf{p} | u_v \rangle$ between the conduction and valence bands. Since we do not know the linear combinations of $u_x, u_y, \& u_z$ that form the valence bands yet, let us look at the momentum-matrix elements $\langle u_s | \mathbf{p} | u_i \rangle$, with $i = x, y, z$. The momentum operator is written out as $\mathbf{p} = -i\hbar(\mathbf{x}\partial/\partial x + \mathbf{y}\partial/\partial y + \mathbf{z}\partial/\partial z)$, and it is immediately clear that

$$\langle u_s | \mathbf{p} | u_i \rangle = \langle u_s | p_i | u_i \rangle \equiv P, \tag{9.11}$$

i.e., it does not vanish. Again, from Figure 2, we can see that the momentum operator along any axis makes the odd-function even, since it is the derivative of that function. The

[1] If the semiconductor has *indirect* bandgap, the conduction-band minimum state is no longer $|S\rangle$-like; it has mixed $|S\rangle$ and p-characteristics.

matrix-element is *defined* to be the constant P. We also note that

$$\langle u_s|p_i|u_j\rangle = 0, (i \neq j). \tag{9.12}$$

To go into a little bit of detail, it can be shown[2] that the valence band states may be written as the following extremely simple linear combinations

$$u_{HH,\uparrow} = -\frac{1}{\sqrt{2}}(u_x + iu_y), \tag{9.13}$$

$$u_{HH,\downarrow} = \frac{1}{\sqrt{2}}(\overline{u_x} - i\overline{u_y}), \tag{9.14}$$

$$u_{LH,\uparrow} = -\frac{1}{\sqrt{6}}(\overline{u_x} + i\overline{u_y} - 2u_z), \tag{9.15}$$

$$u_{LH,\downarrow} = \frac{1}{\sqrt{6}}(u_x - iu_y + 2\overline{u_z}), \tag{9.16}$$

$$u_{SO,\uparrow} = -\frac{1}{\sqrt{3}}(\overline{u_x} + i\overline{u_y} + u_z), \tag{9.17}$$

$$u_{SO,\downarrow} = \frac{1}{\sqrt{3}}(u_x - iu_y - \overline{u_z}) \tag{9.18}$$

and note that

$$\langle u_s|\mathbf{p}|\overline{u_i}\rangle = 0, \tag{9.19}$$

which in words means that Bloch lattice-functions of opposite spins do not interact. With a detailed look at perturbation theory and symmetry properties, we are in the (enviable!) position of understanding $\mathbf{k}\cdot\mathbf{p}$ theory with ease.

9.3 $\mathbf{k}\cdot\mathbf{p}$ theory

Substituting the Bloch wavefunction into Schrodinger equation, we obtain a equation similar to the Schrodinger equation, but with two extra terms -

$$[H^0 + \underbrace{\frac{\hbar}{m_0}\mathbf{k}\cdot\mathbf{p} + \frac{\hbar^2k^2}{2m_0}}_{W}]u(\mathbf{k},\mathbf{r}) = \varepsilon(\mathbf{k})u(\mathbf{k},\mathbf{r}), \tag{9.20}$$

where $u(\mathbf{k},\mathbf{r})$ is the Bloch lattice function.

9.4 No spin-orbit interaction

Let us first look at $\mathbf{k}\cdot\mathbf{p}$ theory *without* spin-orbit interaction. We will return to spin-orbit interaction later. In the absence of spin-orbit interaction, the three valence bands are degenerate at $\mathbf{k} = 0$. Let us denote the bandgap of the (direct-gap) semiconductor by E_g.

Let us look at the eigenvalues at $\mathbf{k} = 0$, i.e., at the Γ point for a direct-gap semiconductor. So the Bloch lattice functions are $u(0,\mathbf{r})$. We assume that we have solved the eigenvalue problem for $\mathbf{k} = 0$, and obtained the various eigenvalues (call then $\varepsilon_n(0)$) for the corresponding eigenstates (call them $|n\rangle$). We look at only four eigenvalues - that of the conduction band ($|c\rangle$) at $\mathbf{k} = 0$, and of the three valence bands - heavy hole ($|HH\rangle$), light hole ($|LH\rangle$) and the split-off band ($|SO\rangle$). In the absence of spin-orbit interaction, they are all degenerate. The corresponding eigenvalues for a cubic crystal are given by ($\varepsilon_c(0) = +E_g, \varepsilon_{HH}(0) = 0, \varepsilon_{LH}(0) = 0, \varepsilon_{SO}(0) = 0$) respectively, where E_g is the (direct) bandgap.

[2] Broido and Sham, Phys. Rev B, **31** 888 (1986)

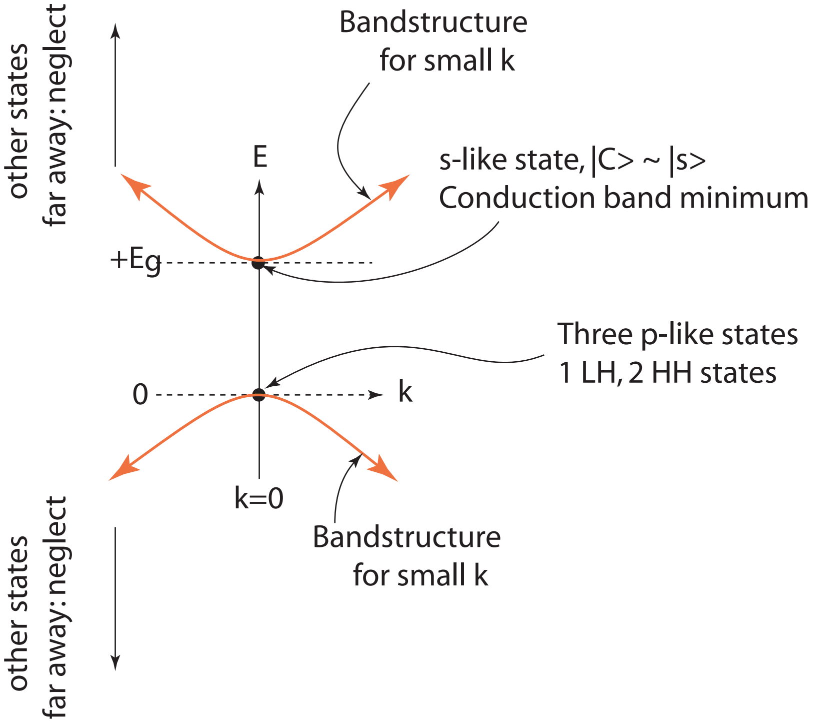


Fig. 9.3: $\mathbf{k}\cdot\mathbf{p}$ bandstructure in the absence of spin-orbit coupling.

Using the two results summarized in the last section, we directly obtain that the n^{th} eigenvalue is perturbed to

$$\varepsilon_n(\mathbf{k}) \approx \varepsilon_n(\mathbf{0}) + \frac{\hbar^2 k^2}{2m_0} + \frac{\hbar^2}{m_0^2}\sum_{m\neq n}\frac{|\langle u_n(0,\mathbf{r})|\mathbf{k}\cdot\mathbf{p}|u_m(0,\mathbf{r})\rangle|^2}{\varepsilon_n(0)-\varepsilon_m(0)}, \tag{9.21}$$

which can be written in a more instructive form as

$$\varepsilon_n(\mathbf{k}) = \varepsilon_n(\mathbf{0}) + \frac{\hbar^2 k^2}{2m^\star}, \tag{9.22}$$

where

$$\frac{1}{m_n^\star} = \frac{1}{m_0}[1 + \frac{2}{m_0 k^2}\sum_{m\neq n}\frac{|\langle u_n(0,\mathbf{r})|\mathbf{k}\cdot\mathbf{p}|u_m(0,\mathbf{r})\rangle|^2}{\varepsilon_n(0)-\varepsilon_m(0)}] \tag{9.23}$$

is the reciprocal *effective mass* of the n^{th} band.

Let us look at the conduction band effective mass. It is given by

$$\frac{1}{m_c^\star} = \frac{1}{m_0}[1 + \frac{2}{m_0 k^2}[\frac{1}{2}(\frac{k^2P^2}{E_g}) + \frac{1}{6}(\frac{k^2P^2}{E_g}) + \frac{1}{3}(\frac{k^2P^2}{E_g})]]. \tag{9.24}$$

Here we have used to form of Bloch lattice functions given in Equations (17)-(22). Cancelling k^2, and recasting the equation, we get

$$m_c^\star \approx \frac{m_0}{1+\frac{2P^2}{m_0 E_G}}. \tag{9.25}$$

To get an estimate of the magnitude of the momentum matrix element P, we do the following. Looking at the momentum matrix element, it is in the form $\langle u_c|\mathbf{p}|u_h\rangle$.

The momentum operator will extract the $k-$value of the state it acts on. Since the valence (and conduction) band edge states actually occur outside the first Brillouin Zone at $|\mathbf{k}| = G = 2\pi/a$ and are folded back in to the Γ-point in the reduced zone scheme, it will extract a value $|P| \approx \hbar \cdot 2\pi/a$, where a is the lattice constant of the crystal. Using this fact, and a typical lattice constant of $a \approx 0.5nm$ we find that

$$\frac{2P^2}{m_0} = \frac{8\pi^2\hbar^2}{m_0 a^2} \approx 24eV. \tag{9.26}$$

In reality, the momentum matrix element of most semiconductors is remarkably constant! In fact, it is a very good approximation to assume that $2P^2/m_0 = 20eV$, which leads to the relation

$$m_c^\star \approx \frac{m_0}{1 + \frac{20eV}{E_G}}, \tag{9.27}$$

which in the limit of narrow-gap semiconductors becomes $m_c^\star \approx (E_g/20)m_0$, bandgap in eV. This is a remarkably simple and powerful result!

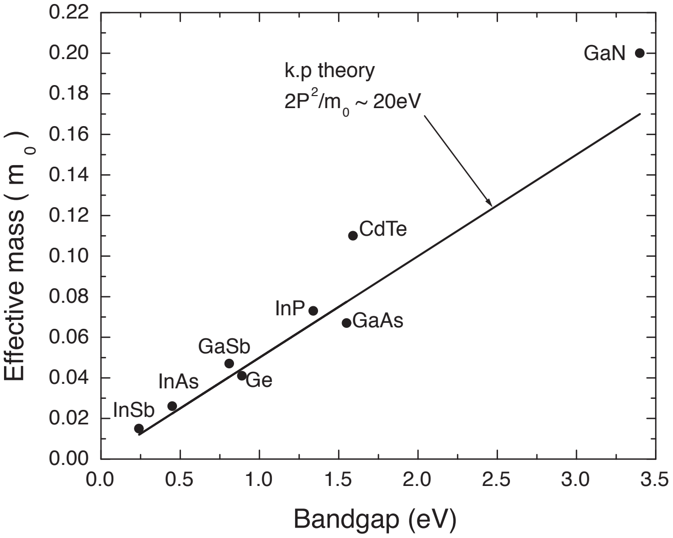


Fig. 9.4: Conduction band effective masses predicted from $\mathbf{k} \cdot \mathbf{p}$ theory. Note that the straight line is an approximate version of the result of $\mathbf{k} \cdot \mathbf{p}$ theory, and it does a rather good job for all semiconductors.

It tells us that the effective mass of electrons in a semiconductor increases as the bandgap increases. We also know exactly *why* this should happen as well: the conduction band energies have the strongest interactions with the valence bands. Since valence band states are lower in energy than the conduction band, they 'push' the energies in the conduction band upwards, increasing the curvature of the band. This directly leads to a lower effective mass. The linear increase of effective mass with bandgap found from the $\mathbf{k} \cdot \mathbf{p}$ theory is plotted in Figure 5 with the experimentally measured conduction band effective masses. One has to concede that theory is rather accurate, and does give a very physical meaning to why the effective mass should scale with the bandgap.

Finally, in the absence of spin-orbit interaction, the bandstructure for the conduction band is

$$\varepsilon_c(\mathbf{k}) \approx E_g + \frac{\hbar^2 k^2}{2m_c^\star}, \tag{9.28}$$

where the conduction band effective mass is used. Note that this result is derived from perturbation theory, and is limited to small regions around the $\mathbf{k} = 0$ point only. One rule of thumb is that the results from this analysis hold only for $|\mathbf{k}| \ll 2\pi/a$, i.e., far from the BZ edges.

9.5 With spin-orbit interaction

What is spin-orbit interaction? First, we have to understand that it is a *purely* relativistic effect (which immediately implies there will be a speed of light c somewhere!). Putting it in words, when electrons move around the positively charged nucleus at relativistic speeds, the electric field of the nucleus Lorentz-transforms to a magnetic field seen by the electrons. The transformation is given by

$$\mathbf{B} = -\frac{1}{2}\frac{(\mathbf{v} \times \mathbf{E})/c^2}{\sqrt{1 - \frac{v^2}{c^2}}} \approx -\frac{1}{2}\frac{\mathbf{v} \times \mathbf{E}}{c^2}, \tag{9.29}$$

where the approximation is for $v \ll c$. To give you an idea, consider a Hydrogen atom - the velocity of electron in the ground state is $v \approx \alpha c$ where $\alpha = \frac{1}{137}$ is the fine structure constant, and the consequent magnetic field seen by such an electron (rotating at a radius $r_0 = 0.53$Å) from the nucleus is - hold your breath - 12 Tesla! That is a *very* large field, and should have perceivable effects.

Spin-orbit splitting occurs in the bandstructure of crystal precisely due to this effect. Specifically, it occurs in semiconductors in the valence band, because the valence electrons are very close to the nucleus, just like electrons around the proton in the hydrogen atom. Furthermore, we can make some predictions about the magnitude of splitting - in general, the splitting should be more for crystals whose constituent atoms have higher atomic number - since the nuclei have more protons, hence more field!

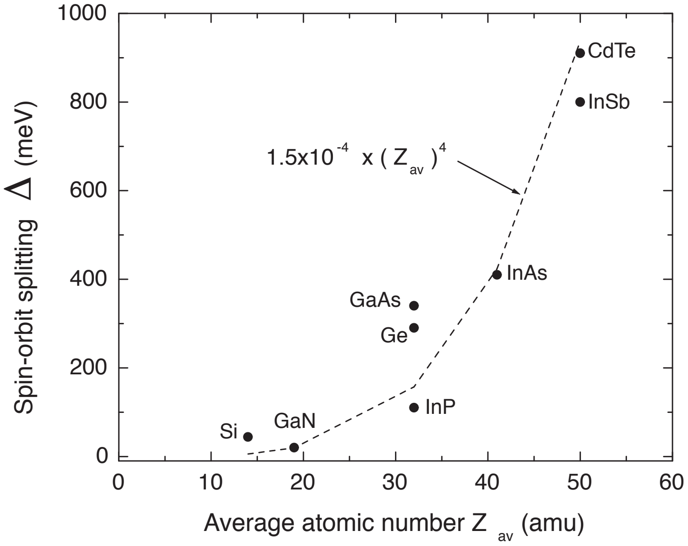


Fig. 9.5: The spin-orbit splitting energy Δ for different semiconductors plotted against the average atomic number Z_{av}. It is a well-known result that the spin-orbit splitting for atomic systems goes as Z^4; the situation is not very different for semiconductors.

In fact, the spin-orbit splitting energy Δ of semiconductors increases as the *fourth* power of the atomic number of the constituent elements. That is because the atomic number is equal to the number of protons, which determines the electric field seen by the valence electrons. I have plotted Δ against average atomic number in Figure 6, and shown a rough fit to a Z_{av}^4 polynomial. For a detailed account on the spin-orbit splitting effects, refer to the textbooks (Yu and Cardona) mentioned in the end of this chapter.

Let us now get back to the business of building in the spin-orbit interaction to the $\mathbf{k} \cdot \mathbf{p}$ theory. Spin-orbit coupling splits the 3 degenerate valence bands at $\mathbf{k} = 0$ into a degenerate HH and LH states, and a split-off state separated by the spin-orbit splitting energy Δ. The eigenvalues at $\mathbf{k} = 0$ are thus given by $(\varepsilon_c(0) = +E_g, \varepsilon_{HH}(0) = 0, \varepsilon_{LH}(0) = 0, \varepsilon_{SO}(0) = -\Delta)$ respectively.

These bandgap E_g, the spin-orbit splitting Δ, and the momentum matrix element P (or, equivalently, the conduction-band effective mass $m_c^\star$) evaluated in the last section are the *inputs* to the $\mathbf{k} \cdot \mathbf{p}$ theory to calculate bandstructure - that is, they are *known*.

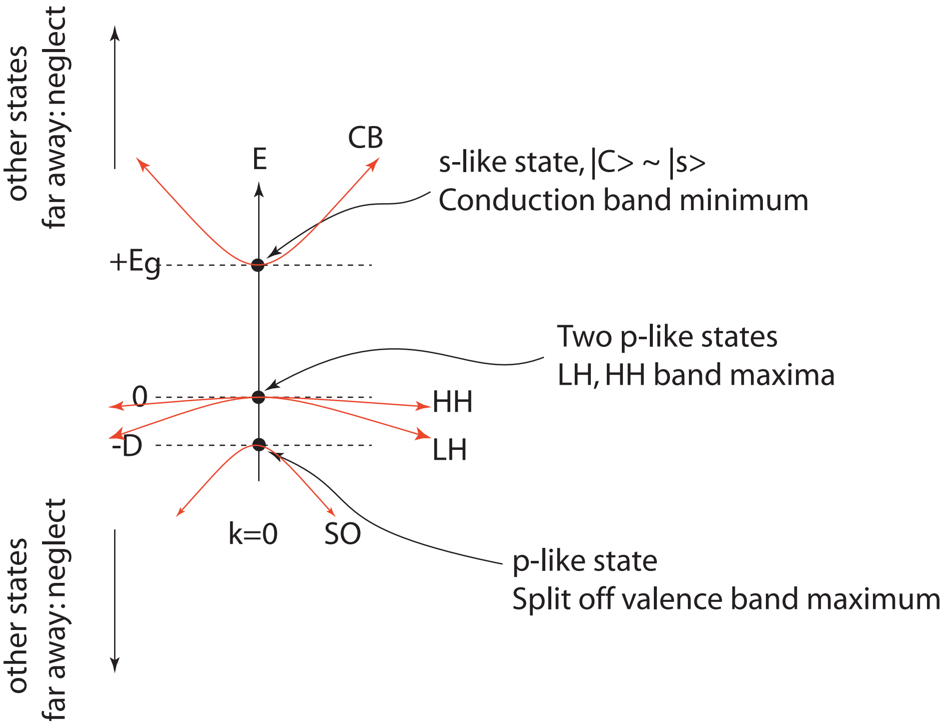


Fig. 9.6: $\mathbf{k} \cdot \mathbf{p}$ bandstructure with spin-orbit splitting.

Using the same results as for the case without spin-orbit splitting, it is rather easy now to show the following. The bandstructure around the Γ point for the four bands and the corresponding effective masses can be written down. For the conduction band, we have

$$\varepsilon_c(k) \approx E_g + \frac{\hbar^2 k^2}{2m_c^\star}, \tag{9.30}$$

where the effective mass is now given by

$$\frac{1}{m_c^\star} = \frac{1}{m_0}[1 + \frac{2}{m_0 k^2}[\frac{1}{2}(\frac{k^2 P^2}{E_g}) + \frac{1}{6}(\frac{k^2 P^2}{E_g}) + \frac{1}{3}(\frac{k^2 P^2}{E_g + \Delta})], \tag{9.31}$$

which can be re-written as

$$m_c^\star \approx \frac{m_0}{1 + \frac{2P^2}{3m_0}(\frac{2}{E_g} + \frac{1}{E_g + \Delta})}, \tag{9.32}$$

which is the same as the case without the SO-splitting if one puts $\Delta = 0$. $2P^2/m_0 \approx 20eV$ is still valid.

Spin-orbit splitting causes changes in the valence bandstructure. We chose not to talk about valence bands in the last section, since the degeneracy prevents us from evaluating the perturbed eigenvalues. However, with spin-orbit splitting, it is easy to show the following.

The HH valence bandstructure is that of a free-electron, i.e., the effective mass is the same as free-electron mass; so,

$$\varepsilon_{HH}(k) = -\frac{\hbar^2 k^2}{2m_0}, \tag{9.33}$$

and the light-hole bandstructure is given by

$$\varepsilon_{LH}(k) = -\frac{\hbar^2 k^2}{2m^\star_{LH}}, \tag{9.34}$$

where the light-hole effective mass is given by

$$m^\star_{LH} = \frac{m_0}{1 + \frac{4P^2}{3m_0 E_g}}. \tag{9.35}$$

Finally, the split-off valence bandstructure is given by

$$\varepsilon_{SO}(k) = -\Delta - \frac{\hbar^2 k^2}{2m^\star_{SO}}, \tag{9.36}$$

where the split-off hole effective mass is given by

$$m^\star_{LH} = \frac{m_0}{1 + \frac{2P^2}{3m_0 (E_g + \Delta)}}. \tag{9.37}$$

This model is known as the Kane-model of $\mathbf{k} \cdot \mathbf{p}$ bandstructure, after Kane's celebrated paper[3] of 1956. There is a very good section on the uses of this form of bandstructure calculation in the text by S. L. Chuang (Physics of Optoelectronic Devices, 1995). $\mathbf{k} \cdot \mathbf{p}$ is very useful in calculating optical transition probabilities and oscillator strengths.

The effects of strain can be incorporated into the $\mathbf{k} \cdot \mathbf{p}$ theory rather easily, and the shifts of bands can be calculated to a great degree of accuracy. The theory is easily extendable to heterostructures, in particular, to quantum wells for calculating density of states, gain in lasers, and so on. The most popular $\mathbf{k} \cdot \mathbf{p}$ calculations employ what is called a 8-band $\mathbf{k} \cdot \mathbf{p}$ formalism. Where do the eight bands come from? We have already seen all 8 - it is the four bands we have been talking about all along, with a spin degeneracy of 2 for each band.

To make the calculations more accurate, one can include bands higher than the conduction band and lower than the valence band. However, the effects of these distant bands are weak, and scale inversely as the energy separation, as we have seen. Thus, they are rarely used.

9.6 Further reading

As Kittel states in his text on Solid State Physics, learning how to calculate bandstructure is an *art*, not learnt from book only, but by experience. My personal favorites for bandstructure theory and applications are two books -

1) *Fundamentals of Semiconductors* (Yu and Cardona, Springer, 1999).
Chapter 2 in this comprehensive text has one of the best modern treatments of semiconductor bandstructure. It makes heavy usage of group theory, which can be intimidating for

[3] E. O. Kane, J. Phys. Chem. Solids, **1**, 82 (1956)

beginners, but nevertheless very rewarding. The authors do not assume that you come all prepared with results from group theory - they actually have ‘crystallized’ the results that are needed from group theory in the chapter.

2) *Energy Bands in Semiconductors* (Donald Long, Interscience Publishers, 1968).
An old and classic monograph, it still remains one of the few books entirely devoted to the topic. The theory is covered in 80 pages, and the rest of the book analyzes bandstructures of specific materials.

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 10

Exact Bandstructure and Going Green

Contents

10.1 Exact Solution: The Kronig-Penney Model 95
10.2 Tight-binding models emerge from Kronig-Penney 99
10.3 Point defects in Kronig-Penney Models 101
10.4 Green's functions and Kronig-Penney for higher-dimensions . . . 103
10.5 Branch points of bandstructure and band-alignments 105

In this chapter, we discuss a non-perturbative or *exactly* solvable model of electron bandstructure in a crystal. It is the celebrated Kronig-Penney model. The purpose of the solution is to illustrate much of bandstructure physics and also to develop a bag of useful concepts that permeate much of solid state physics.

10.1 Exact Solution: The Kronig-Penney Model

An exactly solvable periodic potential problem in quantum mechanics for the electron is the Kronig-Penney model. The problem is exactly solvable in all dimensions - we consider the 1D case. The periodic potential is modeled as a series of Dirac-delta functions

$$V(x) = \sum_n S\delta(x - na), \tag{10.1}$$

where a is the lattice constant, and S is the strength of the perturbation. The sum over n runs over all lattice sites. For example, for a 1D closed ring of length L with $N = L/a$ lattice points and lattice constant a, $0 \leq n \leq N - 1$. This is schematically represented in Figure 10.1.

Now we in the Math primer (**cite**), we have seen the identity $\sum_n \delta(x - na) = \sum_n \frac{1}{a} e^{-i\frac{2\pi}{a}nx}$. Using this relation with $G_n = \frac{2\pi}{a}n$ and substituting in the Schrodinger equation, we get

$$[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{S}{a}\sum_n e^{-iG_n x}]\psi = E\psi. \tag{10.2}$$

The wavefunctions are Bloch functions, which are Fourier expanded in $G_m = \frac{2\pi}{a}m$ as

$$\psi_k(x) = e^{ikx}u_k(x) = e^{ikx}\sum_m u_{G_m}e^{iG_m x} = \sum_m u_{G_m}e^{i(k+G_m)x}, \tag{10.3}$$

Note that $\psi(x = 0) = \sum_{G_m} u_{G_m}$, the sum of *all* the Bloch coefficients, in other words, as long as G's are reciprocal lattice vectors, $\sum_G u_G = \psi(0)$. Now substituting 10.3 in the Schrodinger equation, we get

$$\sum_m \frac{\hbar^2(k + G_m)^2}{2m} u_{G_m}e^{iG_m x} + \frac{S}{a}\sum_m\sum_n u_{G_m}e^{i(G_m - G_n)x} = E_k\sum_m u_{G_m}e^{iG_m x}, \tag{10.4}$$

Multiplying by e^{-iGx} and integrating over all x, we use to identity $\int_0^L e^{i(G'-G)x}dx = L\delta_{G,G'}$ to get

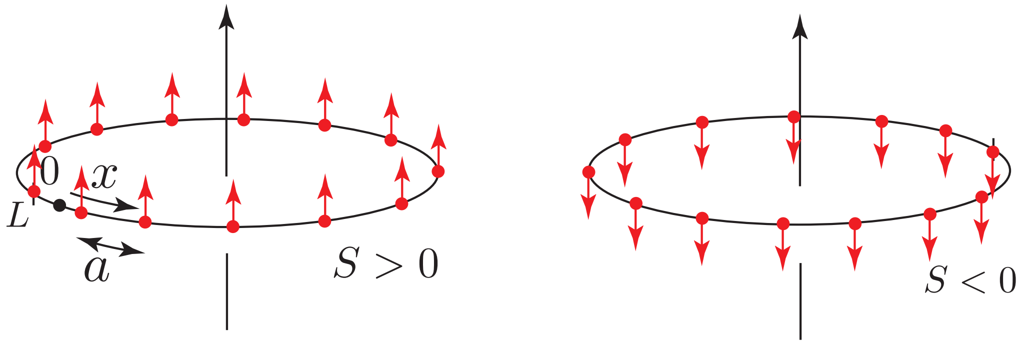


Fig. 10.1: The Kronig-Penney "Dirac" comb periodic potential for a particle on a ring. Left: positive or repulsive potential for $S > 0$, and Right: Negative or attractive potentials for $S < 0$.

$$\frac{\hbar^2(k+G)^2}{2m}u_G + \frac{S}{a}\sum_n u_{G_n+G} = E_k u_G, \tag{10.5}$$

Solving for u_G yields

$$u_G = \frac{S}{a}\frac{\sum_{G_n+G} u_G}{E_k - \frac{\hbar^2(k+G)^2}{2m}}, \tag{10.6}$$

Now for a very useful trick: summing both sides over $G's$ cancels the u_G terms because $\sum_{G_n+G} u_G = \sum_G u_G = \psi(x=0)$, leaving us with the identity

$$\boxed{1 = \frac{S}{a}\sum_G \frac{1}{E_k - \frac{\hbar^2(k+G)^2}{2m}}.} \tag{10.7}$$

This is a rather fancy way of writing unity! Note that this is an *exact* form of the solution of Schrodinger's equation for the periodic potential problem. Inverting it into the form

$$\frac{a}{S} = \sum_G \frac{1}{E_k - \frac{\hbar^2(k+G)^2}{2m}}, \tag{10.8}$$

we are in a position to investigate the aftermath of the solution in Equation 10.7.

Figure 10.2 shows a graphical solution of Equation 10.8 plotted as a function of the energy E_k for two values of k. When the strength of the potential $S > 0$, $a/S > 0$, and is the constant shown in red in the Figure. The RHS is a complex function of energy E_k, with a number of poles located at $E_k = \frac{\hbar^2(k+G)^2}{2m}$, where the RHS diverges. There are several points of intersection - one of which is highlighted. The energies E_k corresponding to these intersection points are the only allowed eigenvalues for the problem. There are several allowed eigenvalues: in fact, there are exactly N distinct eigenvalues corresponding to $n = 0, 1, ..., N-1$ values of $G_n = \frac{2\pi}{a}n$.

If we turn the strength of the potential down by taking $S \to 0$, the red line corresponding to a/S goes off to $+\infty$, and the intersections of the RHS and LHS then are exactly at

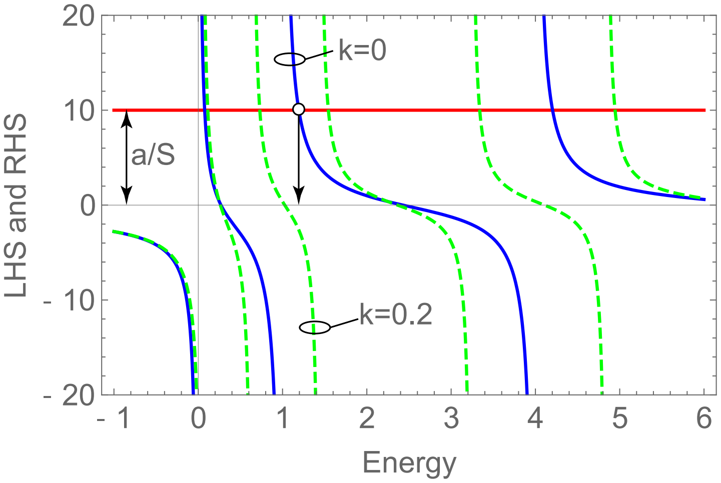


Fig. 10.2: A graphical solution scheme for the repulsive Kronig-Penney Dirac comb. Note that the lowest energy is larger than zero.

the N energies for which the RHS blows up. Clearly, these energy eigenvalues are at $E_k = \frac{\hbar^2(k+G)^2}{2m}$, and we have recovered the nearly free-electron model of the electron.

If the strength is made negative by letting $S < 0$, it is clear that the red line $a/S < 0$, and there is an energy intersection for energy that is *negative*, i.e., $E_k(min) < 0$. This is a "bound" state... or weakly mobile state.

Figure 10.3 shows the calculated energy bandstructures for $S > 0$ (left) and $S < 0$ (right). The axes are in units of $F = \frac{\hbar^2}{2m} \cdot (\frac{\pi}{a})^2$ for energy, and $\frac{2\pi}{a}$ for k. The solid lines in the figure represents several important features of *any* bandstructure in the presence of a non-zero periodic potential. This is superposed on the dashed line plot of bandstructure when the periodic potential is turned off ($S = 0$), but the electron wavefunction is still required to satisfy the lattice periodicity and symmetry, the 'nearly' free-electron (NFE) model with $E = \frac{\hbar^2(k+G)^2}{2m}$. Note that for a repulsive potential with $S > 0$, the Kronig-Penney bandstructure energies are *higher* than the NFE values at *all* values of k except at the Brillouin zone center and edges $k = 0, \pm\frac{\pi}{a}n$. The highest eigenvalue of each Kronig-Penney band is degenerate with the NFE eigenvalues of $E(k = n\frac{\pi}{a}) = n^2 \cdot F$, where $n = 1, 2, ...$, locating energy eigenvalues $F, 9F, ...$ at $k = \pm\frac{\pi}{a}$ at the BZ edge, and $4F, 16F, ...$ at $k = 0$ as the *maxima* of the corresponding bands.

That the energy eigenvalues for $S > 0$ are higher (or equal to) than the NFE values is *guaranteed* by the Hellmann-Feynman theorem. The Hellmann-Feynman theorem states that the eigenvalues E_k of any Hamiltonian $\hat{H}$ satisfy $\frac{\partial E_k}{\partial \lambda} = \langle k|\frac{\partial \hat{H}}{\partial \lambda}|k\rangle$. Imagine the Kronig-Penney potential as a perturbation to the NFE Hamiltonian $\hat{H} = \hat{H}_0 + \lambda\hat{W}$ where $W(x) = S\sum_n \delta(x - na)$, and $\hat{H}_0|k\rangle = E_k^0|n\rangle$, the eigenvalues of the NFE model $E_k^0 = \frac{\hbar^2(k+G)^2}{2m}$ shown by the dashed lines in Figure 10.3. Then, we must have $\frac{\partial E_k}{\partial \lambda} = \langle k|\frac{\partial(\hat{H}_0+\lambda W)}{\partial \lambda}|k\rangle = \langle k|W|k\rangle = \int dx|\psi_k(x)|^2 W(x) = SN|u_k(0)|^2 \geq 0$, and the perturbed eigenvalue $E_k \geq E_k^0$. This remains true at all points in $k-$space except at points of degeneracy, as indicated by an arrow in the left figure of Figure 10.3. At $k-$points were eigenvalues are degenerate, the splitting is such that for $S > 0$, one eigenvalue increases, while the other stays put. The lowest energy allowed is $E_{min}^+ > 0$ for $S > 0$, and the lowest band is rather narrow. This means the electron is 'sluggish' in this band, and it has a large effective mass. As we move up to higher energies, the points of degeneracy develop sharper curvatures and the bands become wider, making the electron effective mass lighter.

Note the differences for the attractive delta potential ($S < 0$) band structures highlighted

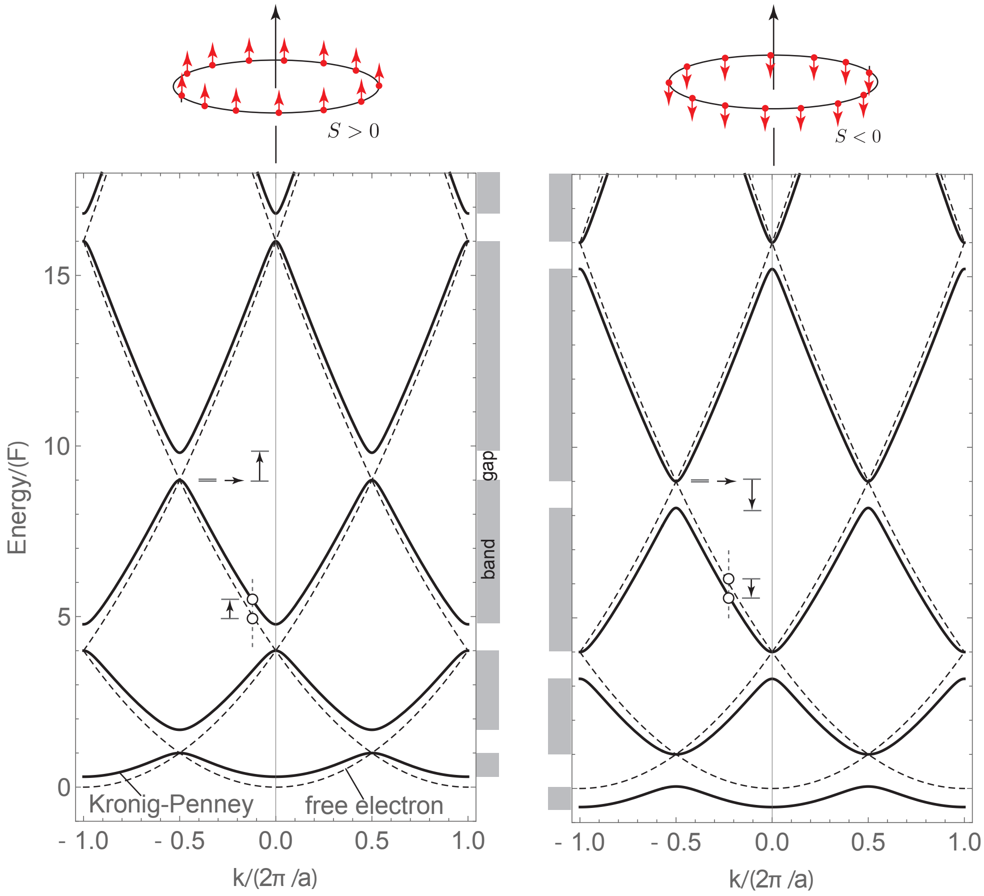


Fig. 10.3: The solid lines show the bandstructure for repulsive (left) and attractive (right) Kronig-Penney potentials. The nearly-free electron bandstructure $E(k) = \frac{\hbar^2(k+G)^2}{2m}$ is shown as dashed lines. The allowed energy bands are indicated in gray along with the energy gaps.

by the right panel in Figure 10.3, and drawn at exactly the same scale for easy comparison. The lowest energy allowed now is $E^{-}_{min} < 0$ for $S < 0$, i.e.. it is *negative* in stark contrast to the situation for $S > 0$. The Hellmann-Feynman theorem now guarantees that the eigenvalues are *lower* than the NFE case. At the $k-$points of degeneracy, the splitting is such that one eigenvalue stays put again, but the other is pushed *down*, exactly opposite to the case of $S > 0$. The lowest eigenvalue of each Kronig-Penney band is degenerate with the NFE eigenvalues of $E(k = n\frac{\pi}{a}) = n^2 \cdot F$ again, where $n = 1, 2, ...$, locating energy eigenvalues $F, 9F, ...$ at $k = \pm\frac{\pi}{a}$ at the BZ edge, and $4F, 16F, ...$ at $k = 0$ as now the *minima* of the corresponding bands.

10.2 Tight-binding models emerge from Kronig-Penney

We will now see that an approximate method to calculate bandstructures called the tight-binding method emerges naturally from the exact Kronig Penney model. Apply the trigonometric identity $\cot(x) = \sum_{-\infty}^{+\infty} \frac{1}{n\pi + x}$ on the right hand side of the central Kronig-Penney eigenvalue equation 10.8, using the fact $G_n = n\frac{2\pi}{a}$. A few trigonometric identities later, equation 10.8 transforms into:

$$\cos(ka) = \cos(qa) + \frac{mSa}{\hbar^2} \cdot \frac{\sin(qa)}{qa}, \tag{10.9}$$

where $q = \sqrt{2mE_k/\hbar^2}$. This is still an *exact* solution of the Schrodinger equation. Now the values of E_k that satisfy this equation will form the energy bandstructure E_k for each k. The left hand side is limited to $-1 \leq \cos(ka) \leq +1$, but the RHS of equation 10.9 can reach values up to $1 + \frac{mSa}{\hbar^2} = 1 + C$ which can clearly exceed unity. This restricts the allowed values of q for real energy eigenvalues $E = \frac{\hbar^2 q^2}{2m}$ for each k. Figure 10.4 shows the 'bands' of q where the RHS lies between $-1 \leq RHS \leq +1$, and real energy eigenvalues are allowed.

Now the zeroes of $\frac{\sin(x)}{x}$ occur at $x = n\pi$ where $n = \pm1, \pm2, ...$ It is clear that a band of $q-$values, and corresponding energies are allowed near the zeroes of the RHS as indicated in Figure 10.4 (left). Let us find an approximate solution for the first band $E_1(k)$ by expanding the RHS for a large strength, or $C = \frac{mSa}{\hbar^2} >> 1$ near the first zero at $n = 1$, around $qa = \pi$. Using $\delta = \pi - qa$, the expansion yields $\cos(qa) + C \cdot \frac{\sin(qa)}{qa} \approx -1 + \frac{C}{\pi}\delta$, which when used in equation 10.9 yields

$$E_1(k) \approx E_0 - 2J(1 + \cos ka), \tag{10.10}$$

where $E_0 = \frac{\pi^2\hbar^2}{2ma^2}$ coincides with the NFE energy at $k = \frac{\pi}{a}$, and the "hopping" or tunneling term is $J = \frac{\pi^2\hbar^4}{2m^2a^3S} = \frac{E_0}{C}$. This is clearly in the form of a tight-binding model! Now we really don't need to stop at the first root - expanding around $qa = n\pi$, and retaining only the linear expansion terms, we get a more comprehensive tight-binding bandstructure of the n^{th} band as:

$$E_n(k) \approx n^2 E_0 \big[1 - \frac{1}{C} + \frac{(-1)^n}{C}\cos(ka)\big]^2. \tag{10.11}$$

Figure 10.4 shows a plot of the first three bands for the dimensionless strength $C = 10$. Note that the energy eigenvalues at the BZ edges co-incide with the free-electron values. This is similar to the case for $S > 0$ in the Kronig Penney model in Figure 10.3 (Left).

Now we can write down a more general tight-binding model by starting from orbitals that are localized at each lattice point, and by trying linear combinations of such orbitals in the Bloch-form to coax out the $E(k)$ eigenvalues and corresponding eigenfunctions. We write the linear combination of atomic orbitals (LCAO) ansatz wavefunction as

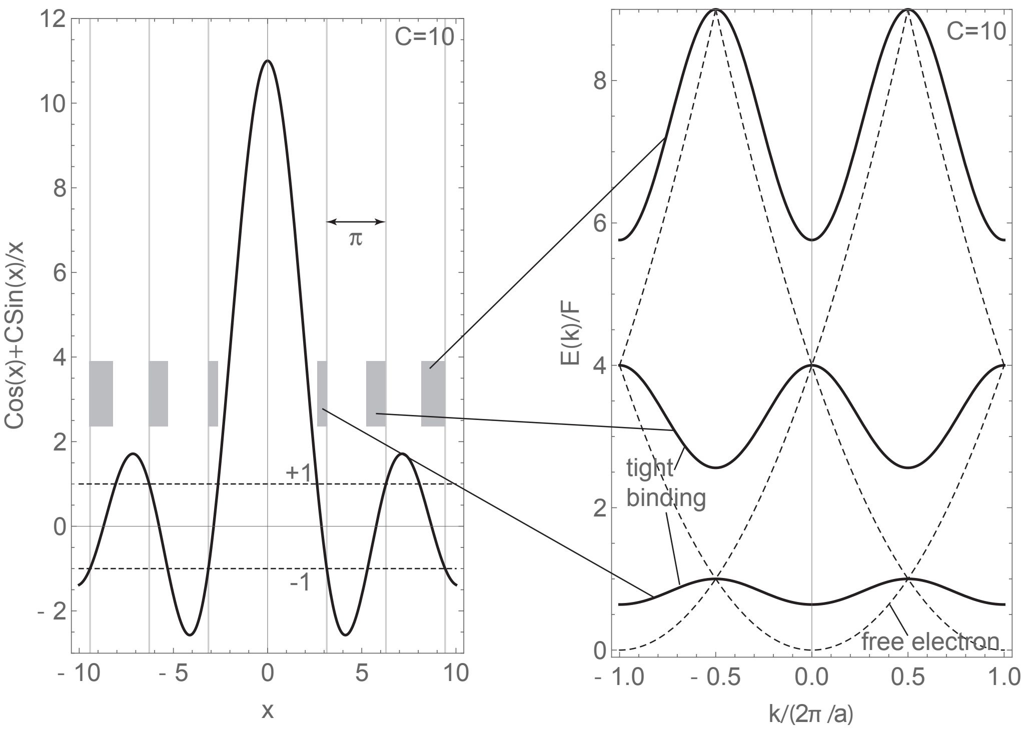


Fig. 10.4: The left figure shows a plot of the RHS of Equation 10.9 with $x = qa$, and the LHS is limited to $-1 \leq LHS \leq +1$. The narrow bands within which the two sides are equal are highlighted; each leads to an allowed energy band. Because the intersections are near $x = n\pi$ where $n = \pm 1, \pm 2, ...$, an approximate analytical expression of all the bands can be obtained (see Equation 10.11). This first three tight-binding bandstructures are plotted in the right panel. Compare with Figure 10.3.

$$|\psi\rangle = \sum_{m=1}^{N} \frac{e^{i\mathbf{k}\cdot\mathbf{R}_m}}{\sqrt{N}}|m\rangle, \tag{10.12}$$

where we have initially assumed just one orbital per lattice site, and $\psi(\mathbf{r}) = \langle\mathbf{r}|\psi\rangle$ and $\phi_m(\mathbf{r}) = \langle\mathbf{r}|m\rangle$ is the orbital centered at site m. This way or writing the ansatz ensures it is indeed a Bloch function, which is verified by checking $\mathbf{r} \to \mathbf{r} + \mathbf{R}$ leads to $\psi(\mathbf{r}+\mathbf{R}) = e^{i\mathbf{k}\cdot\mathbf{R}}\psi(\mathbf{r})$. We feed this ansatz into the Schrodinger equation, cancel $\sqrt{N}$ from each side, and get the relation for energy eigenvalues $E(\mathbf{k})$ for each value of $\mathbf{k}$:

$$\hat{H}\sum_{m=1}^{N} e^{i\mathbf{k}\cdot\mathbf{R}_m}|m\rangle = E(\mathbf{k})\sum_{m=1}^{N} e^{i\mathbf{k}\cdot\mathbf{R}_m}|m\rangle, \tag{10.13}$$

We note that $\hat{H}$ does not affect $e^{i\mathbf{k}\cdot\mathbf{R}_m}$, it acts only on $|m\rangle$. Next we project both vectors on to the ansatz $\sum_{n=1}^{N} e^{i\mathbf{k}\cdot\mathbf{R}_n}\langle n|$, and rearrange to get the central result of the tight-binding energy bandstructure:

$$\boxed{E(\mathbf{k}) = \frac{\sum_{n,m=1}^{N} e^{i\mathbf{k}\cdot(\mathbf{R}_m-\mathbf{R}_n)}\langle n|\hat{H}|m\rangle}{\sum_{n,m=1}^{N} e^{i\mathbf{k}\cdot(\mathbf{R}_m-\mathbf{R}_n)}\langle n|m\rangle}}, \tag{10.14}$$

where ... Now it is clear that there are N^2 terms in the double sum. Out of these, the N 'diagonal' terms are obtained when $n = m$, for which we have $e^{i\mathbf{k}\cdot(\mathbf{R}_m-\mathbf{R}_n)} = e^{i\mathbf{k}\cdot(0)} = 1$ and the diagonal matrix elements are all equal: $\langle n|\hat{H}|n\rangle = E_0$. This energy is slightly *lower* in energy than the original 'atomic' orbital energy because each electron orbital also sees nearby atomic potentials. In the denominator, the diagonal sum gives just $\sum e^{i\mathbf{k}\cdot(0)}\langle n|n\rangle = N$. Let us now look at the rest $N^2 - N$ off-diagonal terms in the numerator and denominator.

Now considering a 1D lattice of lattice constant a, for the 1st nearest neighbors, we have N terms for which the terms in the numerator take the form $\sum_{n=1}^{N}(e^{+ika}\langle n|\hat{H}|n+1\rangle + e^{-ika}\langle n-1|\hat{H}|n\rangle) = -2Nt_1\cos(ka)$, where the hopping integral $t_1 = \langle n|\hat{H}|n+1\rangle$. Similarly, the denominator gives the sum of the N 1st nearest neighbor terms as $+2Ns_1\cos(ka)$, where $s_1 = \langle n|n+1\rangle$ is clearly very small because of decaying wavefunctions that are tightly bound to the lattice sites.

There are N more terms for the 2nd nearest neighbors characterized by the hopping integral $t_2 = \langle n|\hat{H}|n+2\rangle$ and the overlap integral $\langle n|n+2\rangle = s_2$. And then for the 3rd nearest neighbor, and so on... It is intuitively clear that the successive terms t_n and s_n decay fast. Now we can write the expression for the tight-binding bandstructure as:

$$E(k) = \frac{E_0 - 2t_1\cos(ka) - 2t_2\cos(2ka) - 2t_3\cos(3ka)...}{1 + 2s_1\cos(ka) + 2s_2\cos(2ka) + 2s_3\cos(3ka)...} \approx E_0 - 2t_1\cos(ka). \tag{10.15}$$

If instead of 1D, we are in 2D or 3D, then there are more nearest neighbors and the bands acquire more "structure".

10.3 Point defects in Kronig-Penney Models

Now imagine that in the Kronig-Penney model, only one of the N sites has a potential that is different from the other sites. Let us call this difference in the strength U_0, meaning at this particular site, the delta-function strength is $S + U_0$ instead of S, where U_0 can be positive or negative. What is the effect on the energy eigenvalues and the eigenstates due to the presence of this 'defect'?

This problem can now be solved because the exact solution of the Kronig-Penney model *without the defect* has given us the eigenvalues for each $k-$state in the BZ as $E_{KP}(k)$ - for

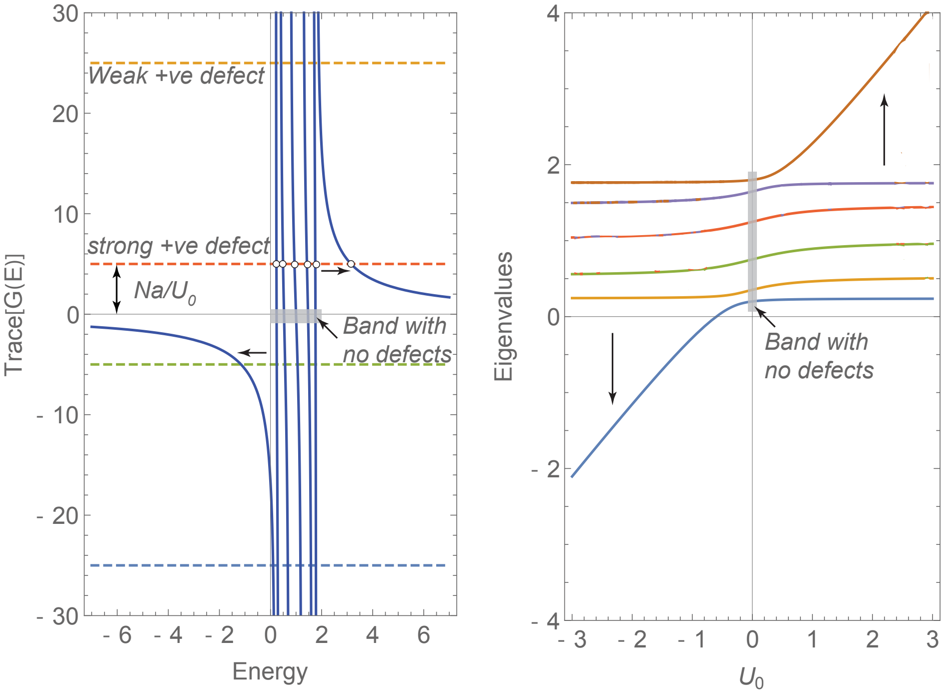


Fig. 10.5: Figures showing the effect of defect states on the allowed energy eigenvalues as a function of the defect potential strength. The left figure shows the graphical solution to the Kronig-Penney type solution, and in particular illustrates the splitting off of one eigenvalue - the highest eigenvalue of the band for positive defect potentials, and the lowest energy eigenvalues for negative defect potentials. This is further highlighted in the figure on the right, where the eigenvalue spectrum is plotted as a function of the defect potential.

example - shown in Figure 10.3. Then, we go through exactly the same procedure that led to the Kronig-Penney solution in Equation 10.8, and end up with the new solution

$$\frac{Na}{U_0} = \sum_k \frac{1}{E_k - E_{KP}(k)}, \tag{10.16}$$

where k are the allowed states in the 1st BZ, N is the number of lattice sites, and therefore $Na = L$ is the macroscopic length. Clearly, in the absence of the defect, $U_0 \to 0$, and the LHS$\to \infty$. This happens exactly N times in the RHS when the allowed energies $E_k = E_{KP}(k)$, i.e., we recover the original Kronig-Penney solution without the defect, as we should.

But when $U_0 \neq 0$, the allowed energies E_k must deviate from $E_{KP}(k)$ to satisfy the exact solution above. To illustrate the solution graphically, we plot the RHS and the LHS in Figure 10.5. We will see in the next section that the RHS of Equations 10.16 and 10.8 are actually the Trace of the Green's function matrix of the problem, i.e., $\sum_k \frac{1}{E_k - E_{KP}(k)} = \text{Trace}[\hat{G}(E)]$. The plot in Figure 10.5 for a few-site chain shows the effect of the defect on the eigenvalue spectrum clearly. The figure on the right illustrates the movement of the eigenvalues as the strength of the defect is tuned from zero to large positive and large negative. The eigenvalues at $U_0 = 0$ constitute the band without the defect. When the defect strength is +ve and strong, the LHS L/U_0 line moves closer to the $x-$axis (left figure), and it is clear that one of the intersections - at the *top* of the energy band splits off from the band rapidly, whereas all other eigenvalues do not change as much. Any change is positive, as guaranteed by the Hellmann-Feynman theorem. This is a characteristic feature - similarly, for a negative U_0, the lowest eigenvalue of the band splits off and leaves other eigenvalues mostly unchanged.

We will see later that $U_0 > 0$ 'defects' explain the formation of acceptor states at the top of valence bands, and are designed such that the splitting energy is less than kT for room-temperature generation of holes. Similarly, the bottom of the band with $U_0 < 0$ models donor states and electron doping at the bottom of the conduction band of semiconductors.

10.4 Green's functions and Kronig-Penney for higher-dimensions

We noted the repeated appearance of sums over the Brillouin zone of the kind $\sum_k \frac{1}{E-E(k)}$ which have units of (energy)$^{-1}$. This may be thought of as a function of the variable E, or energy. The reason why such sums permeate exact solutions of problems will now become clear: and will lead us to define Green's functions.

Consider the Schrodinger equation

$$i\hbar\frac{\partial}{\partial t}\Psi = \hat{H}\Psi \to [i\hbar\frac{\partial}{\partial t} - \hat{H}]\Psi = 0. \tag{10.17}$$

Let us think of the equation as the product of the operator (or matrix) $i\hbar\frac{\partial}{\partial t} - \hat{H}$ with Ψ. For this product to be zero, either $i\hbar\frac{\partial}{\partial t} - \hat{H}$ or Ψ, or both should be zero. The only interesting case here is when we actually have a quantum object with a nonzero wavefunction, $\Psi \neq 0$. Thus, $i\hbar\frac{\partial}{\partial t} - \hat{H}$ should be zero. Now we have learnt that if the quantum object is in a state of definite energy, $i\hbar\frac{\partial}{\partial t}\Psi_n = E_n\Psi_n$, Ψ_n, and E_n is a real eigenvalue representing the energy of the state. Let us generalize this and write $i\hbar\frac{\partial}{\partial t} = E$, where E is a *variable*. We can then write the Schrodinger equation as $[EI - \hat{H}]\Psi = 0$, where I is an identity operator, or the identity matrix when the equation is written out for any chosen basis. However, the equation in this form does not hold true for all E, but only for certain $E = E_n$ - only when the variable E matches up with an allowed eigenvalue.

Now let us think of $EI - \hat{H}$ as a *function* of E. When we vary E, this function has very sharp responses when $E = E_n$: the function is a 'detector' of eigenvalues - it detects an eigenvalue by vanishing. At those sharp energies, $\Psi = \Psi_n \neq 0$ is an eigenfunction, so the function provides the eigenfunction as its 'residue'. Now with this qualitative picture in mind, let us solidify the concept of the Green's function of the system.

We like detectors to 'scream' when they detect, rather than to go silent. So, can we find a function $\hat{G}$ that instead of solving the equation $[EI - \hat{H}]\Psi = 0$, solves the equation $[EI - \hat{H}]\hat{G} = I$ instead? Formally, the function is clearly $\hat{G} = [EI - \hat{H}]^{-1}$. This function clearly blows up when $E = E_n$, and is indeed the screaming detector we are looking for. It is the Green's function for the Hamiltonian $\hat{H}$. Let us assume that we know all the eigenvalues of a particular Hamiltonian $\hat{H}_0$ to be E_n and the corresponding eigenfunctions are $|n\rangle$. The Green's function can then be written out as a matrix form

$$\hat{G}_0(E) = \sum_n [EI - \hat{H}]^{-1}|n\rangle\langle n| = \sum_n \frac{|n\rangle\langle n|}{E - E_n}. \tag{10.18}$$

It is clear that the Green's function is actually a matrix, and sums of the kind that appeared earlier in the solution of the Kronig-Penney and the defect problems are the sum of the diagonal terms in a diagonal basis. Now it turns out that the sum of the diagonal terms is *invariant* with what basis one writes the matrix - which is why it goes by a name - the Trace. Thus, we have a very important relation

$$\text{Trace}[\hat{G}(E)] = \sum_k \frac{1}{E - E_0(k)} \tag{10.19}$$

where $E_0(k)$ are the allowed eigenvalues of the system. The solution of the Kronig-Penney model is thus very compactly written in the formal way as $\text{Trace}[\hat{G}_0(E)] = \frac{a}{S}$, where $\hat{G}_0(E) = (EI - \hat{H}_0)^{-1}$, and $\hat{H}_0|k\rangle = E_0(k)|k\rangle$ are the nearly-free electron eigensystem, with $E_0(k) = \frac{\hbar^2(k+G)^2}{2m}$. The solution of a single-site defect state of strength S_0 is then written as $\text{Trace}[\hat{G}(E)] = \frac{Na}{S_0}$, where now the Green's function is for the exactly solved Kronig-Penney eigensystem $\text{Trace}[\hat{G}(E)] = (EI - \hat{H})^{-1}$, where $\hat{H}|k\rangle = E_{KP}(k)|k\rangle$, and $E_{KP}(k)$ are the Kronig-Penney eigenvalues.

More on Green's functions - relation to DOS, etc to be written...

We can write the Green's function in a non-diagonal basis as well. For example, we can write instead,

$$\hat{G}_0(l, l'; E) = \sum_n \langle l[EI - \hat{H}]^{-1}|n\rangle\langle n|l'\rangle = \sum_n \frac{\langle l|n\rangle\langle n|l'\rangle}{E - E_n}, \tag{10.20}$$

which includes both diagonal and off-diagonal terms of the matrix. Since this form does not require us to start with a diagonal basis, it is therefore preferred.

At this point we make the connection with the Kronig-Penney model by identifying the eigenvalue index n as the allowed $k-$points in the Brillouin zone - there are N of them. We have also written the Wannier wavefunction at site l' as $\langle n|l'\rangle = \sum_k \frac{e^{ik\cdot R_{l'}}}{\sqrt{N}}\phi_0(x)$. So the Green's function can then be written as

$$\hat{G}_0(l, l'; E) = \sum_k \frac{e^{ik(R_l - R_{l'})}}{E - E(k)} = \int_{-\frac{\pi}{a}}^{+\frac{\pi}{a}} \frac{dk}{\frac{2\pi}{L}} \frac{e^{ik(R_l - R_{l'})}}{E - E(k)}, \tag{10.21}$$

Following text till the left arrows is work in progress → →.

10.5 Branch points of bandstructure and band-alignments

There exists an unique energy value E_{BP} for any bandstructure, for which we have the condition

$$\sum_{\mathbf{k}}[E_{BP} - E_{lower}(\mathbf{k})] = \sum_{\mathbf{k}}[E_{higher}(\mathbf{k}) - E_{BP}] \tag{10.22}$$

which in some sense is the 'weighted average' energy level. Let there be n_+ bands with higher energy than E_{BP} and n_- bands with lower energy. The sum over $\mathbf{k}$ runs over the entire Brillouin zone, say over N states, which is equal to the macroscopic number of lattice points in the crystal. Then, we have

$$N \cdot n_- \cdot E_{BP} - \sum_{\mathbf{k},i} E_{lower,i}(\mathbf{k}) = \sum_{\mathbf{k},j} E_{higher,j}(\mathbf{k}) - N \cdot n_+ \cdot E_{BP}, \tag{10.23}$$

which gives us the Branch point at

$$E_{BP} = \frac{\sum_{\mathbf{k},i} E_{lower,i}(\mathbf{k}) + \sum_{\mathbf{k},j} E_{higher,j}(\mathbf{k})}{N(n_- + n_+)}, \tag{10.24}$$

← ← Preceeding text till the right arrows is work in progress

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 11

Doping and Semiconductor Heterostructures: The Effective Mass Method

Contents

11.1 Effective Mass Approximation, Envelope Functions 107
11.2 3D, 2D, 2D, 0D: Heterostructures 110
11.3 Bulk Bandstructure . . . 111
11.4 Doped Semiconductors . . 112
11.5 Quantum Wells 113
11.6 Quantum Wires 114
11.7 Quantum Dots 116

11.1 Effective Mass Approximation, Envelope Functions

Before we jump into considering real semiconductors with impurities and corresponding perturbations from perfect periodic potentials, it is worthwhile to develop a very powerful formalism that greatly simplifies our treatment of transport properties. So long as the perturbations of the crystal potential is not drastic, one can re-cast the Schrödinger equation in a form that is very useful for discussing transport and device applications. One runs into a fundamental problem in dealing with a particle location in real space and its momentum at the same time. To do that, the concept of a wave packet is necessary. Wave packets, unlike pure Bloch-eigenstates, have a finite spread both in the momentum and real space. A wave packet is nothing but a linear combination of Bloch eigenstates for small $k-$values around a region of interest in the Brillouin zone. For most cases, it suffices to investigate properties of electrons and holes located very close to the band extrema in the $k-$space; therefore, one collects Bloch eigenstates around such points, and creates a wavepacket by taking their linear combinations.

To illustrate this, let us consider the 1-dimensional case. We construct a wavepacket by taking a linear combination of Bloch eigenstates $\phi_{nk}(x)$ from the n^{th} band with wavevector k. The sum is over the whole BZ.

$$\psi(x) = \sum_n \sum_k C(k)\phi_{nk}(x) = \sum_n \int \frac{dk}{2\pi} C(k)\phi_{nk}(x) \tag{11.1}$$

We now make two crucial approximation -
a) We assume that wavefunctions from only one band play a part in the wavepacket, and thus drop the sum over all bands.
b) We assume that in the single band we are interested in, wavevectors from a small region say around $k_0 = 0$ are important (see Figure 11.1).

Then, Bloch functions can be written as $\phi_{nk}(x) = e^{ikx}u_{nk}(x) \approx u_{n0}(x)e^{ikx} = \phi_{n0}(x)e^{ikx}$. Then the wavepacket takes the form

$$\psi(x) \approx \phi_{n0}(x) \int \frac{dk}{2\pi} C(k)e^{ikx} = \underbrace{\phi_{n0}(x)}_{\text{Bloch}} \cdot \underbrace{C(x)}_{\text{envelope}}, \tag{11.2}$$

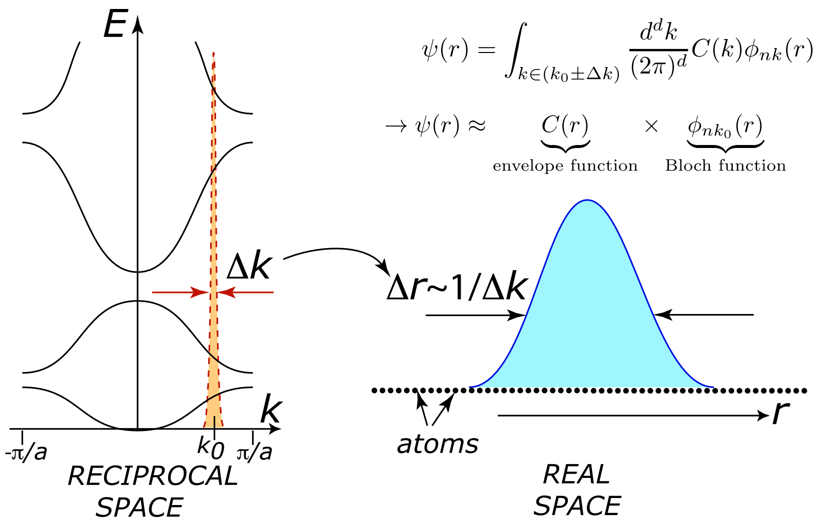

E
$\psi(r) = \int_{k\in(k_0\pm\Delta k)} \frac{d^dk}{(2\pi)^d} C(k)\phi_{nk}(r)$
$\rightarrow \psi(r) \approx \underbrace{C(r)}_{\text{envelope function}} \times \underbrace{\phi_{nk_0}(r)}_{\text{Bloch function}}$
Δk
Δr~1/Δk
k
-π/a
k0
π/a
RECIPROCAL SPACE
atoms
r
REAL SPACE

Fig. 11.1: A wavepacket is constructed by taking Bloch functions from a small region of the reciprocal space, and summing them with weights. The weights $C(k)$ have a small extent Δk in reciprocal space; when carried over to real space, the spread is large, since $\Delta r \sim 1/\Delta k$; thus the wavepacket has a finite spread in real space, and represents the wavefunction of a particle. If we restrict the sum in reciprocal space to 1% of the BZ, the wavepacket spreads over $1/0.01 = 100$ atoms in real space. The real space wavefunction is given by the Bloch wavefunction at the k_0 point, modulated by an envelope function $C(r)$, which is the Fourier transform of the weights $C(k)$.

where the integral term is identified as the Fourier transform of the weights $C(k) \leftrightarrow C(x)$. The real-space function $C(x)$ which is a Fourier transform of the weights of the wavepacket is called as the **envelope** function; since the weights $C(k)$ are over a small region in $k-$space, $C(x)$ is spread over real space. It is typically a smooth function spreading over several lattice constants. This is illustrated in Figure 11.2.

How does the wavepacket behave when we apply the periodic crystal Hamiltonian H_0 on it? Since $\phi_{nk}(x)$ are Bloch-eigenfunctions of this Hamiltonian, $H_0\phi_{nk}(x) = E_n(k)\phi_{nk}(x)$, and we recover

$$H_0\psi(x) = \int \frac{dk}{2\pi} C(k)E_n(k)\phi_{nk}(x) \approx \phi_{n0}(x)\int \frac{dk}{2\pi} C(k)E_n(k)e^{ikx}. \tag{11.3}$$

We now write out the energy eigenvalues as a Taylor-series of small wavevectors around $k = k_0 = 0$,

$$E_n(k) = \sum_m a_m k^m \tag{11.4}$$

and Schrödinger equation becomes

$$H_0\psi(x) \approx \phi_{n0}(x)\sum_m \int \frac{dk}{2\pi} C(k)a_m k^m e^{ikx}. \tag{11.5}$$

We now use a property of Fourier transforms - if $f(k) \leftrightarrow f(x)$, then $kf(k) \leftrightarrow (-id/dx)f(x)$, and in general, $k^m f(k) \leftrightarrow (-id/dx)^m f(x)$. Thus,

$$\int \frac{dk}{2\pi} k^m C(k)e^{ikx} \leftrightarrow (-i\frac{d}{dx})^m C(x), \tag{11.6}$$

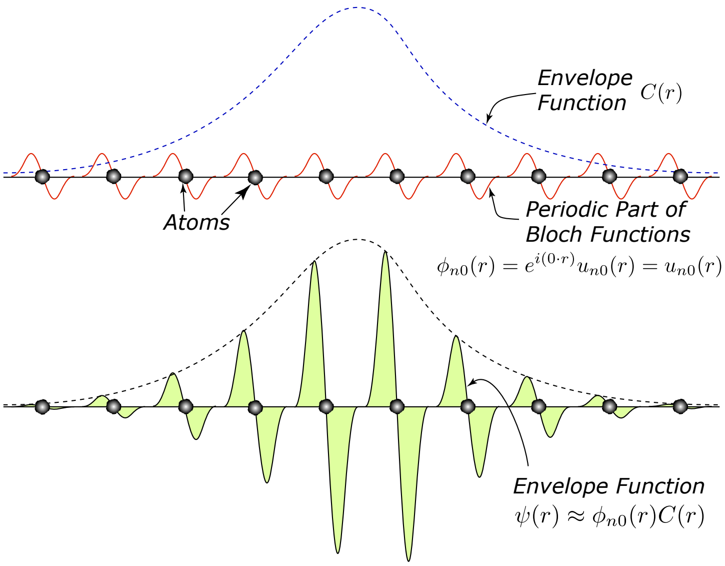

Envelope Function $C(r)$
Atoms
Periodic Part of Bloch Functions
$\phi_{n0}(r) = e^{i(0\cdot r)}u_{n0}(r) = u_{n0}(r)$
Envelope Function
$\psi(r) \approx \phi_{n0}(r)C(r)$

Fig. 11.2: Envelope function $C(r)$ modulates the Bloch function $\phi_{n0}(x)$ to produce the wavefunction of the wavepacket $\psi(x)$.

and the Schrödinger equation is recast as

$$H_0\psi(x) \approx \phi_{n0}(x)E_n(-i\nabla)C(x), \tag{11.7}$$

which can be generalized to the 3-D case. Thus, in the energy term, we make the substitution $k \to i\partial/\partial r$, making it an operator that acts on the envelope function only. This step is crucial - the Bloch function part has been pulled out as a coefficient; no operators act on it.

Now, instead of the periodic potential Hamiltonian, if we have another potential (say a perturbation) $W(r)$ present, Schrödinger equation becomes

$$H_0\phi_{n0}(r)C(r) + W(r)\phi_{n0}(r)C(r) = E\phi_{n0}(r)C(r), \tag{11.8}$$

and using Equation 11.7, it becomes

$$[E_n(-i\nabla) + W(r)]C(r) = EC(r), \tag{11.9}$$

where the Bloch functions do not appear at all! Furthermore, if we already know the bandstructure of the semiconductor, then we can write the energy around the point $k_0 = 0$ of interest in terms of the effective mass, and the operator $E_n(-i\nabla)$ thus becomes

$$E_n(k) \approx E_c(r) + \frac{\hbar^2k^2}{2m^\star} \to E_n(-i\nabla) \approx E_c(r) - \frac{\hbar^2}{2m^\star}\nabla^2, \tag{11.10}$$

and the Schrödinger equation takes the enormously simplified form

$$[-\frac{\hbar^2}{2m^\star}\nabla^2 + W(r)]C(r) = [E - E_c(r)]C(r), \tag{11.11}$$

which is the celebrated “Effective Mass Approximation”. Take a moment to note what has been achieved. The Schrodinger equation has been re-cast into a much simpler problem of a particle of mass $m^\star$, moving in a potential $E_c(r) + W(r)$! All information about the bandstructure and crystal potential has been lumped into the effective mass $m^\star$. The wavefunctions are envelope functions $C(r)$, from which one recovers the real wavefunction of the wavepacket by multiplying with the Bloch function - $\psi(r) \approx \phi_{n0}(r)C(r) = u_{n0}(r)C(r)$. The envelope functions $C(r)$ can be actually determined for any potential - it amounts to solving the Schrödinger equation for a particle in the potential $E_c(r) + W(r)$. Note that the envelope function in the absence of any impurity potential $W(r) = 0$ is given by

$$C(r) = \frac{1}{\sqrt{V}} e^{i\mathbf{k}\cdot\mathbf{r}}, \tag{11.12}$$

and the corresponding eigenvalues of the Schroödinger equation are given by

$$E = E_c(r) + \frac{\hbar^2 |\mathbf{k}|^2}{2m^\star}. \tag{11.13}$$

If we consider electrons at the bottom of the conduction band, $E_c(r)$ is the spatial variation of the conduction band edge - exactly what one draws in band diagrams. An impurity potential can now be included as a perturbation to the periodic crystal, and the new energy eigenvalues can be found. As an example, consider an ionized impurity, which has a Coulomb potential. The effective mass equation takes the form

$$[-\frac{\hbar^2}{2m^\star}\nabla^2 - \frac{e^2}{4\pi\epsilon r}]C(r) = (E - E_c)C(r), \tag{11.14}$$

which is identified as the same as the classic problem of a hydrogen atom, albeit with two modifications - the mass term is an effective mass instead of the free electron mass, and the dielectric constant is that of the semiconductor. Then, the new energy levels that appear are given by

$$E - E_c = E_\infty \frac{m^\star}{\epsilon_r^2}, \tag{11.15}$$

and the effective Bohr-radius is given by

$$a_B^\star = a_B \frac{\epsilon_r}{m^\star} \tag{11.16}$$

In bulk semiconductors, the band-edge variation in real space can be varied by applying electric fields, or by doping variations. In semiconductor heterostructures, one can further engineer the variation of the band-edge $E_c(r)$ in space by quasi-electric fields - the band edge can behave as quantum-wells, wires, or dots, depending upon composition of the semiconductor. The effective mass approximation is a natural point of departure, where analysis of such low-dimensional structures begins.

11.2 3D, 2D, 2D, 0D: Heterostructures

With the explosion of usage of semiconductor heterostructures in a variety of applications, low-dimensional structures such a quantum wells, wires and dots have become important. They come in various avatars - these structures can be grown by compositional variations in epitaxially grown semiconductor layers by MBE/MOCVD techniques, or nanowires / nanotubes / nanocrystals can be grown by bottom-up approaches (by CVD techniques, or by solution chemistry). So, understanding bandstructure of these artificially engineered materials is of great interest.

The goal of many clever expitaxial/bottom-up techniques to create nanostructures amounts to modifying the bandstructure of the constituent bulk semiconductor material. Many of these designer materials have niche applications, and have a potential to perform functions that are difficult, if not impossible to achieve in bulk materials. An example is the semiconductor (diode) laser. The first semiconductor lasers were band-engineering by doping (i.e., they were p-n junctions).

We have derived the effective mass equation for carriers in bulk semiconductors in the envelope-function approximation. The three-dimensional effective mass equation is

$$[-\frac{\hbar^2}{2m^\star}\nabla^2 + V(r)]C(r) = (E - E_c(r))C(r). \tag{11.17}$$

Here, $C(r)$ is the envelope function of carriers in the band under consideration. The Schrödinger equation is thus re-cast in a form which is identical to that of an electron in a total potential $V(r) + E_c(r)$, determined by the band-edge behavior. It has mapped the complex problem of an electron moving through a crystal experiencing very complicated potentials to a textbook-type 'particle in a well-defined potential' problem, which is solvable. The particle mass is renormalized, absorbing the details of the crystal potential. The *real* wavefunction of the wavepacket that models the particle-like nature of the electrons is given by $\psi(r) \approx u_{n0}(r)C(r)$, where $u_{n0}(r)$ is the periodic part of the Bloch eigenstates of the crystal that result from the periodic crystal potential. However, the beauty of the effective mass approximation is that the envelope function is all that is needed to find the bandstructure of the low-dimensional structures[1]! The envelope function concept is a powerful tool, as is demonstrated in its use in determining bandstructure modifications due to quantum confinement of carriers in low-dimensional structures.

11.3 Bulk Bandstructure

In a bulk semiconductor in the absence of external fields, $V(r) + E_c(r) = E_{c0}$ is a constant energy (flatband conditions), and thus the solution of the effective mass equation yields envelope functions

$$C(r) = \frac{1}{\sqrt{V}} e^{i\vec{k}\cdot\vec{r}}, \tag{11.18}$$

and energies

$$E(k) = E_{c0}(r) + \frac{\hbar^2 k^2}{2m^\star} = E_{c0}(r) + \frac{\hbar^2}{2}(\frac{k_x^2}{m_{xx}^\star} + \frac{k_y^2}{m_{yy}^\star} + \frac{k_z^2}{m_{zz}^\star}). \tag{11.19}$$

One should not forget that even thought the k^s is written as a continuous variable, they are actually quantized, assuming values

$$k_x = k_y = k_z = \frac{2\pi}{L} m \tag{11.20}$$

where $m = 0, \pm 1, \pm 2, ...$ Since L is a macroscopic length, the quantization is very fine, and for all practical purposes, k^s can be assumed continuous.

The density of states (DOS) is given by

$$g_{3D}(E) = \frac{1}{2\pi^2}(\frac{2m^\star}{\hbar^2})^{3/2}\sqrt{E - E_{c0}}, \tag{11.21}$$

from which one gets a carrier concentration in the conduction band

$$n = \int_0^\infty dE f_{FD}(E) g_{3D}(E) = N_C^{3D} F_{1/2}(\frac{E_C - E_F}{k_B T}) \approx N_C^{3D} e^{-\frac{E_C - E_F}{k_B T}}, \tag{11.22}$$

[1] Note that the bulk bandstructure is assumed to be known. The effective mass contains information about the bulk bandstructure.

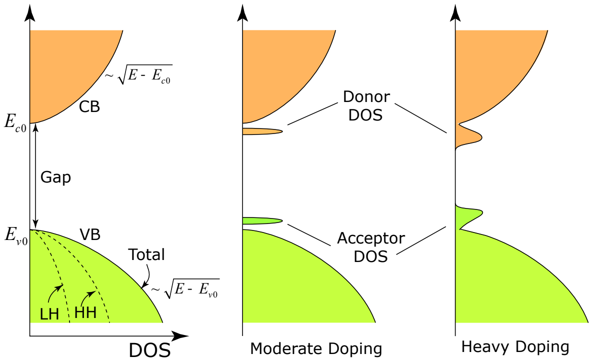


Fig. 11.3: Density of States of bulk (undoped), moderately doped and heavily doped semiconductors.

where $F_j(...)$ is the Fermi-Dirac integral function. The approximation holds only when Fermi-Dirac distribution can be approximated by a Maxwell-Boltzmann form. Here, it is easily shown that N_C^{3D} is a effective band-edge DOS is

$$N_C^{3D} = 2(\frac{m^\star k_B T}{2\pi\hbar^2})^{\frac{3}{2}}. \tag{11.23}$$

Similar results hold for valence bands, where the contributions from the Light and Heavy hole bands add to give the total DOS. This is shown schematically in Figure 11.3.

11.4 Doped Semiconductors

Doping adds states in the bandgap of the semiconductor. A shallow dopant adds states close to the band-edges. Considering a shallow donor, the Hydrogenic-model solution from the effective mass equation

$$[-\frac{\hbar^2}{2m^\star}\nabla^2 - \frac{e^2}{4\pi\epsilon r}]C(r) = (E - E_c)C(r) \tag{11.24}$$

showed that the eigenvalues were similar to that of a hydrogen atom, given by $E_n = E_{c0} - Ry^\star/n^2$, where $Ry^\star = 13.6 \times (m^\star)/\epsilon_r^2$ is the modified Hydrogenic energy levels. The ground-state envelope functions around the donor atoms

$$C(r) \sim e^{-r/r_0} \tag{11.25}$$

is spread over many lattice constants ($r_0 = a_B(\epsilon_r/m^\star) \gg a$); this implies that in k-space, the donor states are localized to $\Delta k \sim 1/r$. If the donor electron envelope function is spread over 1000 atoms in real space, in k-space it will be restricted to ∼1/1000 of the volume of the Brilloiun zone. Thus, for all practical purposes, the donor states are assumed to be "atomic-like". Energy separations between these individual atomic-like states is very small.

For heavy doping however, many changes can occur. The adjacent radii of electrons associated with adjacent donors can overlap, leading to formation of impurity bands. Then, the semiconductor acquires metal-like properties, since thermal activation of carriers into the bands is not necessary for transport. The effects of moderate and heavy doping on the DOS of bulk semiconductors is shown in Figure 11.3.

11.5 Quantum Wells

Quantum wells are formed upon sandwiching a thin layer of semiconductor between wider bandgap barrier layers. The finite extent of the quantum well layer makes the conduction band profile mimic a one-dimensional quantum well in the direction of growth ($z-$direction), leaving motion in the $x-y$ plane free. Thus, the square-well potential (with reference to the conduction band edge E_{c0}) is written as

$$V(x,y,z) = 0, z < 0 \tag{11.26}$$

$$V(x,y,z) = 0, z > W \tag{11.27}$$

$$V(x,y,z) = -\Delta E_c, 0 \leq z \leq W. \tag{11.28}$$

Using the effective mass equation with this potential, it is evident that the envelope function should decompose as

$$C_{n_z}(x,y,z) = \phi(x,y)\chi_{n_z}(z) = [\frac{1}{\sqrt{A}}e^{i(k_x x + k_y y)}] \cdot [\chi_{n_z}(z)] \tag{11.29}$$

If the quantum well is assumed to be infinitely deep, by simple wave-fitting procedure[2] the $z-$component of the electron quasi-momentum is quantized to

$$k_{n_z} = \frac{\pi}{W}n_z, \tag{11.30}$$

where $n_z = 1, 2, 3, \ldots$. From simple particle-in-a-box model in quantum mechanics, the normalized $z-$component of the envelope function is

$$\chi_{n_z}(z) = \frac{2}{\sqrt{W}} \sin\frac{\pi n_z z}{W}. \tag{11.31}$$

The bandstructure is the set of energy eigenvalues is obtained from the effective mass equation, given by

$$E(k) = E_{c0} + \underbrace{\frac{\hbar^2}{2}(\frac{k_x^2}{m_{xx}^\star} + \frac{k_y^2}{m_{yy}^\star})}_{E_{2D}(k_x,k_y} + \underbrace{\frac{\hbar^2}{2m_{zz}^\star}(\frac{\pi n_z}{W})^2}_{E_{1D}(n_z)} \tag{11.32}$$

which evidently decomposes to a free-electron component in the $x-y$ plane and a quantized component in the $z-$ direction. The bandstructure consists of multiple bands $E_{2D}(k_x, k_y)$, each indexed by the quantum number n_z; this is shown in Figure 11.4.

The DOS of electrons confined in an ideal 2-D plane is a constant, given by $g_{2D}(E) = m^\star/\pi\hbar^2$. In the quantum well, each subband corresponding to an n_z is an ideal 2-D system, and each subband contributes $g_{2D}(E)$ the the total DOS. This is shown schematically in Figure 11.4. Thus, the DOS of the quantum well is

$$g_{QW}(E) = \frac{m^\star}{\pi\hbar^2}\sum_{n_z}\theta(E - E_{n_z}), \tag{11.33}$$

where $\theta(\ldots)$ is the unit step function. The carrier density of an ideal 2D electron system is thus given by

$$n_{2D} = \int_0^\infty dE f_{FD}(E) g_{2D}(E) = \underbrace{\frac{m^\star k_B T}{\pi\hbar^2}}_{N_C^{2D}} \ln(1 + e^{\frac{E_F - E_1}{k_B T}}), \tag{11.34}$$

[2] Only waves that satisfy $n_z(\lambda/2) = W$ fit into the well of width W, leading to $k_{n_z} = 2\pi/\lambda = (\pi/W)n_z$.

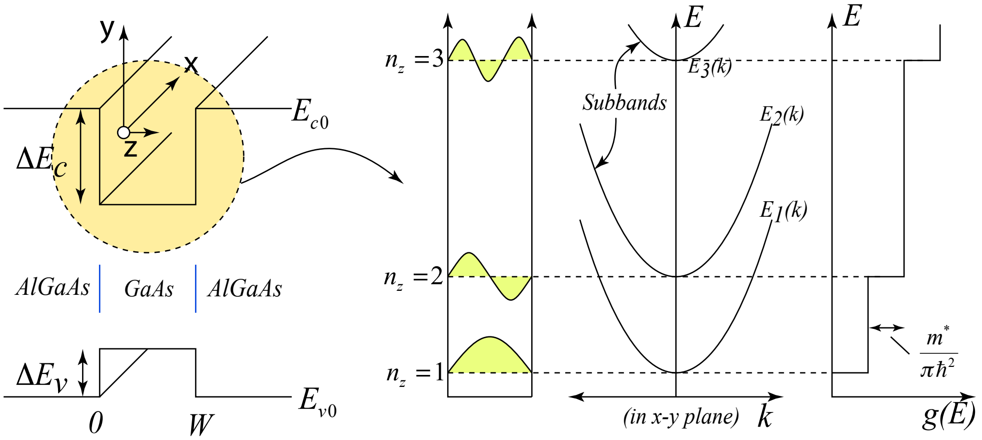


Fig. 11.4: Bandstructure, and DOS of realistic heterostructure quantum wells.

where E_1 is the ground state energy, E_F is the Fermi level, and N_C^{2D} is the effective band-edge DOS, the 2-dimensional counterpart of N_C^{3D} defined in Equation 11.23. (Verify the units of each!)

For the quantum well, which houses many subbands, the DOS becomes a sum of each subband (Figure 11.4), and the total carrier density is thus a sum of 2D-carriers housed in each subband -

$$n_{2D} = \sum_j n_j = N_c^{2D} \sum_j \ln(1 + e^{\frac{E_F - E_j}{k_B T}}). \tag{11.35}$$

Note that for a 2D system, no approximation of the Fermi-Dirac function is necessary to find the carrier density analytically.

It is important to note that if the confining potential in the $z-$direction can be engineered almost at will by modern epitaxial techniques by controlling the spatial changes in material composition. For example, a popular quantum well structure has a parabolic potential ($V(z) \sim z^2$), which leads to the E_{n_z} values spaced in equal energy intervals - this is a characteristic of a square, or Harmonic Oscillator potential. Another extremely important quantum well structure is the triangular well potential ($V(z) \sim z$), which appears in MOSFETs, HEMTs, and quantum wells under electric fields. The triangular well leads to E_{n_z} values given by Airy funtions. Regardless of these details specific to the *shape* of the potential, the bandstructure and the DOS remain similar to the square well case; the only modification being the E_{n_z} values, and the corresponding subband separations.

11.6 Quantum Wires

Artificial quantum wires are formed either lithographically (top-down approach), or by direct growth in the form of semiconductor nanowires or nanotubes (bottom-up approach). In a quantum well, out of the three degrees of freedom for real space motion, carriers were confined in one, and were free to move in the other two. In a quantum wire, electrons are free to move freely in one dimension only (hence the name 'wire'), and the other two degrees of freedom are quantum-confined. Assume that the length of the wire (total length L_z) is along the $z-$direction (see Figure 11.5), and the wire is quantum-confined in the $x - y$ plane ($L_x, L_y \ll L_z$). Then, the envelope function naturally decomposes into

$$C(x, y, z) = \chi_{n_x}(x) \cdot \chi_{n_y}(y) \cdot (\frac{1}{\sqrt{L_z}} e^{ik_x x}), \tag{11.36}$$

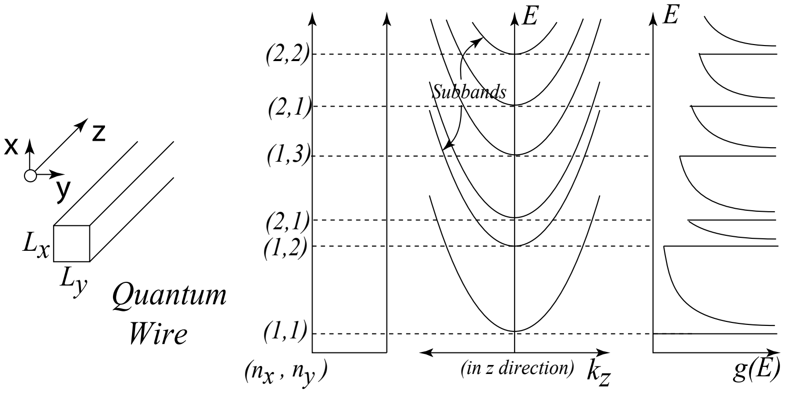


Fig. 11.5: Bandstructure, and DOS of realistic quantum wires.

and the energy eigenvalues are given by

$$E(n_x, n_y, k_z) = E(n_x, n_y) + \frac{\hbar^2 k_k^2}{2m^\star_{zz}}. \tag{11.37}$$

If the confinement in the $x - y$ directions is by infinite potentials (a useful model applicable in many quantum wires), then similar to the quantum well situation, a wave-fitting procedure gives

$$k_{n_x} = \frac{\pi}{L_x} n_x, \tag{11.38}$$

$$k_{n_y} = \frac{\pi}{L_y} n_y, \tag{11.39}$$

where $n_x, n_y = 1, 2, 3, \ldots$ independently.
The eigenfunctions assume the form

$$C_{n_x,n_y}(x, y, z) = [\sqrt{\frac{2}{L_x}} \sin(\frac{\pi n_x}{L_x} x)] \cdot [\sqrt{\frac{2}{L_y}} \sin(\frac{\pi n_y}{L_y} y)] \cdot [\frac{1}{\sqrt{L_z}} e^{ik_x x}], \tag{11.40}$$

and the corresponding bandstructure is given by

$$E(n_x, n_y, k_z) = \underbrace{[\frac{\hbar^2}{2m_{xx}} (\frac{\pi n_x}{L_x})^2] + [\frac{\hbar^2}{2m_{yy}} (\frac{\pi n_y}{L_y})^2]}_{E(n_x,n_y)} + \frac{\hbar^2 k_z^2}{2m^\star_{zz}}. \tag{11.41}$$

Multiple subbands are formed, similar to the quantum well structure. A new subband forms at each eigenvalue $E(n_x, n_y)$, and each subband has a dispersion $E(k_z) = \hbar^2 k_z^2 / 2m_{zz}$ (Figure 11.5).

The DOS of electrons confined to an ideal 1-D potential is given by

$$g_{1D}(E) = \frac{1}{\pi}\sqrt{\frac{2m^{\star}}{\hbar^2}}\frac{1}{\sqrt{E-E_1}}, \tag{11.42}$$

where E_1 is the lowest allowed energy (ground state). Due to multiple subbands, the DOS acquires peaks at every eigenvalue $E(n_x, n_y)$. Since there are two quantum numbers involved, some eigenvalues can be degenerate, and the peaks can occur at irregular intervals as opposed to the quantum well case. The general DOS for a quantum wire can thus be written as

$$g_{QWire}(E) = \frac{1}{\pi}\sqrt{\frac{2m^{\star}}{\hbar^2}}\sum_{n_x,n_y}\frac{1}{\sqrt{E-E(n_x,n_y)}}, \tag{11.43}$$

which is shown schematically in Figure 11.5.

11.7 Quantum Dots

The quantum dot is the ultimate nanostructure. All three degrees of freedom are quantum confined; therefore there is no plane-wave component of electron wavefunctions. The envelope function for a "quantum box" of sides L_x, L_y, L_z (see Figure 11.6) is thus written as

$$C(x,y,z) = \chi_{n_x}(x)\chi_{n_y}(y)\chi_{n_z}(z), \tag{11.44}$$

and if the confining potential is infinitely strong, we have $k_{n_i} = (\pi/L_i)n_i$ for $i = x, y, z$. The envelope functions are thus given by

$$C(x,y,z) = [\sqrt{\frac{2}{L_x}}\sin(\frac{\pi n_x}{L_x})]\cdot[\sqrt{\frac{2}{L_y}}\sin(\frac{\pi n_y}{L_y})]\cdot[\sqrt{\frac{2}{L_z}}\sin(\frac{\pi n_z}{L_z})], \tag{11.45}$$

and the energy eigenvalues are given by

$$E(n_x,n_y,n_z) = \frac{\hbar^2}{2m_{xx}}(\frac{\pi n_x}{L_x})^2 + \frac{\hbar^2}{2m_{yy}}(\frac{\pi n_y}{L_y})^2 + \frac{\hbar^2}{2m_{zz}}(\frac{\pi n_z}{L_z})^2. \tag{11.46}$$

Note that the the energy eigenvalues are no more quasi-continuous, and are indexed by three quantum numbers (n_x, n_y, n_z). Thus, it does not make sense to talk about "bandstructure" of quantum dots; the DOS is a sum of delta functions, written as

$$g_{QDot} = \sum_{n_x,n_y,n_z}\delta(E - E_{n_x,n_y,n_z}). \tag{11.47}$$

This is shown schematically in Figure 11.6. Since there is no direction of free motion, there is no transport *within* a quantum dot, and there is no quasi-continuous momentum components. Fabricating quantum dots by lithographic techniques is pushing the limits of top-down approach to the problem. On the other hand, epitaxial techniques can coax quantum dots to self-assemble by cleverly exploiting the strain in lattice-mismatched semiconductors. On the other hand, bottom-up techniques of growing nanocrystals in solution by chemical synthetic routes is becoming increasingly popular.

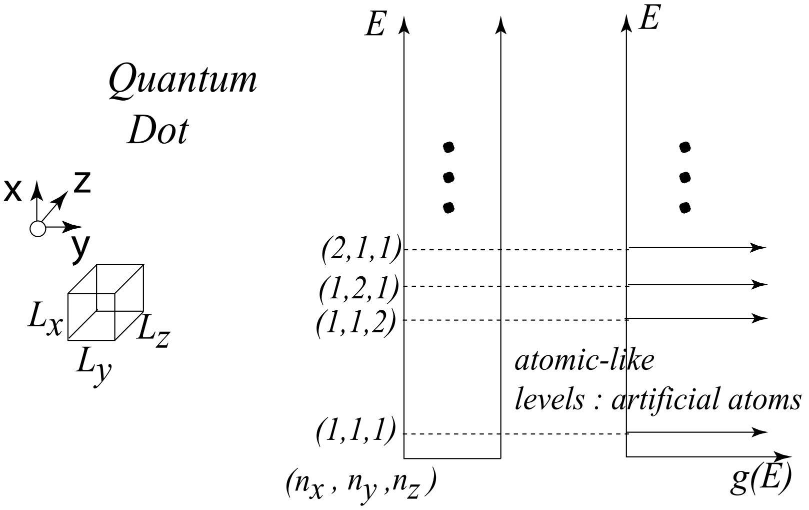


Fig. 11.6: Energy levels and DOS of quantum dots.

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 12

The Ballistic Transistor

Contents

12.1 Introduction 119
12.2 The field-effect transistor 119
12.3 Ballistic current-voltage characteristics 123
12.4 Examples 125

12.1 Introduction

In this chapter, we apply the formalism we have developed for charge currents to understand the output characteristics of a field-effect transistor. Specifically, we consider the situation when transport of electrons in the transistor occurs without scattering due to defects, i.e., ballistically from the source contact to the drain. The ballistic characteristics highlight various quantum limits of performance of a transistor. They guide material and geometry choices to extract the most of such devices. In this process we develop powerful insights into the inner workings of the remarkable device that powers the digital world.

12.2 The field-effect transistor

Figure 12.1 illustrates a typical field-effect transistor. A 2-dimensional electron gas (2DEG) at the surface of a semiconductor (or in a quantum well) is the conducting channel. It is separated from a gate metal by a barrier of thickness t_b and dielectric constant ϵ_b. The gate metal electrostatically controls the 2DEG density via the capacitance $C_b = \epsilon_b/t_b$. The source and the drain metals form low-resistance ohmic contacts to heavily doped regions indicated in gray. The FET width in the y-direction is W, which is much larger than the source-drain separation L and the barrier thickness t_b.

The 2DEG density at different points x of the channel from the source to the drain depends on the relative strength of the electrostatic control of the three contacts. We assume that the source contact is grounded. V_{ds} is the drain potential and V_{gs} is the gate potential with respect to the source. When $V_{ds} = 0$ V, the 2DEG forms the lower plate of a parallel-plate capacitor with the gate metal. A threshold voltage V_T is necessary on the gate to create the 2DEG. Once created, the 2D charge density n_s in the 2DEG changes as $qn_s \approx C_g(V_{gs} - V_T)$, where $C_g = C_b C_q/(C_b + C_q)$, where C_q is a density-of-states or 'quantum' capacitance. Note that $qn_s \approx C_g(V_{gs} - V_T)$ is true only in the 'on-state' of the transistor, and will not give us the sub-threshold or off-state characteristics. The quantum capacitance arises because the density of states of the semiconductor band is lower than the metal: this forces a finite voltage drop in the semiconductor to hold charge. It may also be pictured as a finite spread of the 2DEG electrons, whose centroid is located away from the surface, adding an extra capacitance in series to the barrier capacitance. We will use the zero-temperature limit of $C_q \approx q^2 \times \rho_{2d}$ for our purposes here, where $\rho_{2d} = g_s g_v m^\star/2\pi\hbar^2$ is the DOS for each subband of the 2DEG. Since $V_{ds} = 0$ V, no *net* current flows from the source to the drain. However, when the 2DEG is present, the electrons are carrying current. The microscopic picture is best understood in the $\mathbf{k}-$space.

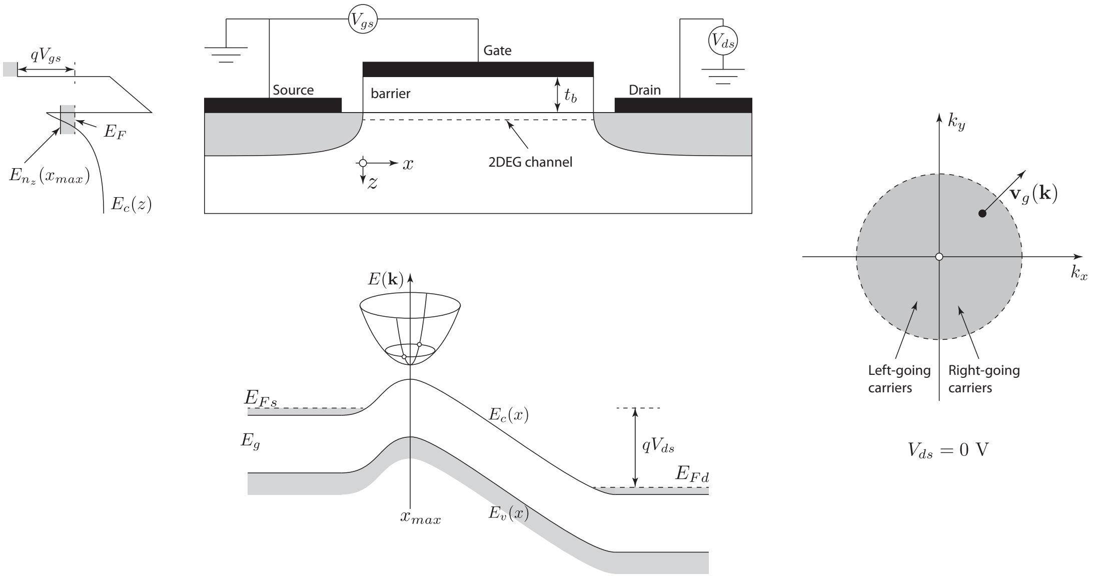


Fig. 12.1: Field effect transistor, energy band diagram, and $\mathbf{k}-$space occupation of states.

The states of the first subband of the 2DEG are illustrated in the real-space energy band diagram and the occupation picture in $\mathbf{k}-$space in Figure 12.1. When $V_{gs} > V_T$, a quantum-well is created with the $z-$quantization resulting in a ground state energy E_{n_z}. The total energy of electrons in this 2DEG subband is given by

$$E(k_x, k_y) = E_c + E_{n_z} + \frac{\hbar^2(k_x^2 + k_y^2)}{2m^\star}, \tag{12.1}$$

where E_c is the conduction band edge energy at the interface, and $m^\star$ is the effective mass of the sub-bandstructure. We choose $E_c = 0$, and $m^\star$ to be isotropic. When $V_{ds} = 0$ V, the 2DEG electrons are in equilibrium with the source and drain. So the Fermi-level of the 2DEG electrons E_F is the same as the source and the drain. The band edge E_c and quantization energy E_{n_z} have to adjust to populate the channel with the charge dictated by the gate capacitor $qn_s = C_g(V_{gs} - V_T)$. The Fermi-Dirac distribution dictates the carrier distribution of the 2DEG in the $\mathbf{k}-$space. It is given by

$$f(k_x, k_y) = \frac{1}{1 + \exp\left[\left(\frac{\hbar^2}{2m^\star}(k_x^2 + k_y^2) - (E_F - E_{n_z})\right)/kT\right]} = \frac{1}{1 + \exp\left[\frac{\hbar^2(k_x^2+k_y^2)}{2m^\star kT} - \eta\right]}, \tag{12.2}$$

where we define $\eta = (E_F - E_{n_z})/kT$. Since the Fermi-level is controlled by the gate alone when $V_{ds} = 0$, we should be able to write η as a function of the gate voltage V_{gs}. The relation comes about by summing all occupied states in the $\mathbf{k}-$space:

$$C_g(V_{gs} - V_T) = q \underbrace{\frac{g_s g_v}{LW} \int \frac{dk_x}{\frac{2\pi}{L}} \frac{dk_y}{\frac{2\pi}{W}} \frac{1}{1 + \exp\left[\frac{\hbar^2(k_x^2+k_y^2)}{2m^\star kT} - \eta\right]}}_{n_s} = q \frac{g_s g_v}{(2\pi)^2} \int_0^\infty \int_0^{2\pi} \frac{k dk d\theta}{1 + \exp\left[\frac{\hbar^2 k^2}{2m^\star kT} - \eta\right]}. \tag{12.3}$$

We made the substitution $k_x = k\cos\theta$ and $k_y = k\sin\theta$. Pictorially, we are summing the states, or finding the 'area' of occupied states in the $\mathbf{k}-$space in Figure 12.1. At zero temperature, the shape is a circle with a sharp edge indicated by the dashed circle. At higher temperatures, the edge is diffuse, and the occupation probability drops exponentially as it is crossed. The spin-degeneracy of each state is g_s, and the semiconductor has g_v equivalent valleys, each with the same bandstructure.

The integral in Equation 12.3 is evaluated by first integrating out over θ which gives a factor 2π, and then making the substitution $u = \hbar^2k^2/2m^\star kT$. Doing so with $V_{th} = kT/q$ yields

$$C_g(V_{gs} - V_T) = q\frac{g_sg_vm^\star kT}{2\pi\hbar^2}\underbrace{\int_0^\infty \frac{du}{1+\exp\left[u-\eta\right]}}_{F_0(\eta)} = C_qV_{th}F_0(\eta), \tag{12.4}$$

where we identify $C_q \approx q^2\rho_{2d} = q^2g_sg_vm^\star/2\pi\hbar^2$ as the quantum capacitance, and the integral $F_0(\eta)$ as a special case of generalized Fermi-Dirac integrals of the form

$$F_j(\eta) = \int_0^\infty du\frac{u^j}{1+\exp\left[u-\eta\right]}, \tag{12.5}$$

with $j = 0$. The zeroth order Fermi-Dirac integral evaluates exactly to $F_0(\eta) = \ln[1+\exp(\eta)]$. At this stage, it is useful to define $\eta_g = \frac{C_b}{C_b+C_q}(\frac{V_{gs}-V_T}{V_{th}})$. Thus the gate voltage V_{gs} tunes the Fermi level E_F of the 2DEG according to the relation

$$\eta = \frac{E_F - E_{n_z}}{kT} = \ln\left(e^{\eta_g} - 1\right). \tag{12.6}$$

For $V_{gs} - V_T >> V_{th}$, $\eta_g >> 1$, and we obtain $\eta \approx \eta_g$, implying $E_F - E_{n_z} \approx q(V_{gs} - V_T) \times C_b/(C_b + C_q)$. In other words, at a high gate overdrive voltage, the Fermi level changes approximately linearly with the gate voltage, as one would expect in a parallel plate capacitor. The capacitance factor is less than one, indicating a voltage division between the barrier and the channel. A part of the voltage must be spent to create the 2DEG since the density of states of the semiconductor conduction band is much smaller than a metal, as is apparent from the energy band diagram along the $z-$direction in Figure 12.1.

If we are interested in evaluating the sub-threshold characteristics of the ballistic FET, Equation 12.4 must be modified. It is evident that the RHS of this equation is always +ve, but when $V_{gs} < V_T$ in the sub-threshold, the LHS is -ve. To fix this problem, by looking at the energy band diagram in Figure 12.1 we rewrite the division of voltage drops as $qV_b + (E_F - E_{n_z}) = q(V_{gs} - V_T)$, where V_T now absorbs the surface barrier height, the conduction band offset between the barrier and the semiconductor, and the ground state quantization energy $(E_{n_z} - E_c)$. The term V_b is the voltage drop in the barrier given by $V_b = F_bt_b = (qn_s/\epsilon_b)t_b = qn_s/C_b$. The resulting relation between n_s and V_{gs} is then

$$\frac{q^2n_s}{C_b} + kT\ln\left(e^{\frac{qn_s}{C_qV_{th}}} - 1\right) = q(V_{gs} - V_T) \implies \boxed{e^{\frac{qn_s}{C_bV_{th}}}\left(e^{\frac{qn_s}{C_qV_{th}}} - 1\right) = e^{\frac{V_{gs}-V_T}{V_{th}}}} \tag{12.7}$$

This is a transcendental equation, which must be numerically solved to obtain n_s as a function of V_{gs} to get the functional dependence $n_s(V_{gs})$. Note that since $n_s > 0$, both sides of the equation *always* remain +ve. As $V_{gs} - V_T$ becomes large and negative, $n_s \to 0$ exponentially but never reaches 0. This is the sub threshold characteristics of the ballistic transistor. In Equation 12.7, two characteristic carrier densities appear: $n_b = C_bV_{th}/q$ and $n_q = C_qV_{th}/q$; the equation then reads $e^{\frac{n_s}{n_b}}(e^{\frac{n_s}{n_q}} - 1) = e^{\frac{V_{gs}-V_T}{V_{th}}}$. For $V_{gs} - V_T >> V_{th}$, the 1 in the bracket may be neglected, and $qn_s \approx \frac{C_bC_q}{C_b+C_q}(V_{gs} - V_T)$. On the other hand, when $V_{gs} - V_T << 0$, the RHS is small. Since $n_s > 0$, it must become very small. Expanding the exponentials and retaining the leading order, we obtain $n_s \approx n_q e^{\frac{V_{gs}-V_T}{V_{th}}}$. In the sub threshold regime, the carrier density at the source-injection point decreases *exponentially* with the gate voltage, and is responsible for the sharp switching of the device. Figure 12.2 illustrates this behavior. For the rest of the chapter, we focus on the on-state of the ballistic FET.

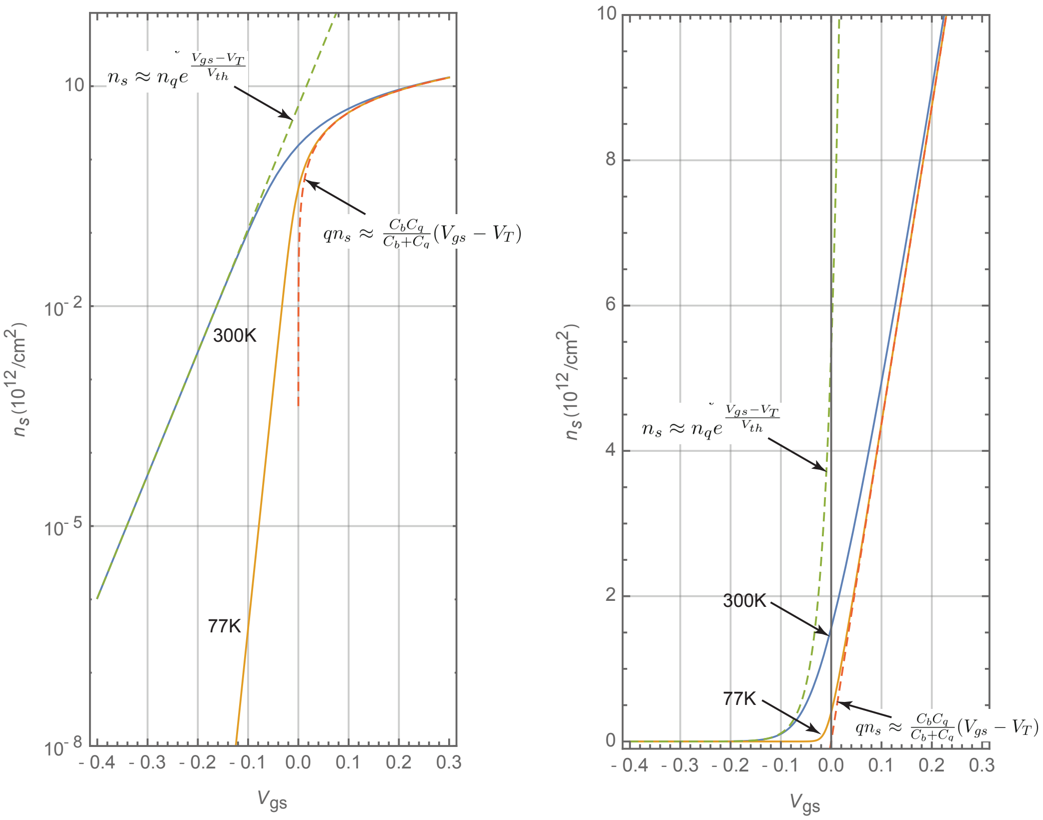


Fig. 12.2: Illustrating the dependence of the 2DEG sheet density at the injection point on the gate voltage.

At this stage, it is instructive to find the right-going and left-going components of the net current at $V_{ds} = 0$ V, even though the net current is zero. We derived the general quantum-mechanical expression for current flowing in $d-$dimensions earlier as

$$\mathbf{J}_d = \frac{q g_s g_v}{(2\pi)^d} \int d^d\mathbf{k} \times \mathbf{v}_g(\mathbf{k}) f(\mathbf{k}), \tag{12.8}$$

where we assumed the transmission probability $T(\mathbf{k}) = 1$. For the 2DEG here, $d = 2$ and the group velocity of state $|\mathbf{k}\rangle$ is $\mathbf{v}_g(\mathbf{k}) = \hbar\mathbf{k}/m^\star$. From Figure 12.1, this velocity component points radially outwards from the origin in $\mathbf{k}-$space. Clearly evaluating this integral will yield zero since there is a $|-\mathbf{k}\rangle$ state corresponding to each $|+\mathbf{k}\rangle$ state. So instead, we evaluate the current carried by electrons moving *only* in the $+k_x = |\mathbf{k}|\cos\theta = k\cos\theta$ direction. This is obtained from Eq. 12.8 by restricting the $\mathbf{k}-$space integral to the right half plane covered by $-\pi/2 \leq \theta \leq +\pi/2$ and using the velocity projected along the k_x axis $v_g = \hbar k\cos\theta/m^\star$ to obtain

$$J_{2d}^{\rightarrow} = \frac{q g_s g_v \hbar}{(2\pi)^2 m^\star} \int_{k=0}^{\infty} \int_{\theta=-\frac{\pi}{2}}^{+\frac{\pi}{2}} \frac{(k\cos\theta)kdkd\theta}{1+\exp\left[\frac{\hbar^2k^2}{2m^\star kT}-\eta\right]} = \underbrace{\frac{q g_s g_v \sqrt{2m^\star}(kT)^{\frac{3}{2}}}{2\pi^2\hbar^2}}_{J_0^{2d}} F_{\frac{1}{2}}(\eta), \tag{12.9}$$

where $F_{1/2}(\eta)$ is the dimensionless Fermi-Dirac integral of order $j = 1/2$, and the prefactor J_0^{2d} has units of A/m or current per unit width. Since $J_{2d}^{\rightarrow} = J_{2d}^{\leftarrow} = J_0^{2d} F_{1/2}(\eta)$, the net current is zero. Another way to visualize this is to think of the right-going carriers as being created by injection into the 2DEG channel from the source, and thus the right-half carriers in $\mathbf{k}-$space are in equilibrium with the source. This statement is quantified by requiring $E_F^{\rightarrow} = E_{Fs}$. Similarly, the left-going carriers are injected from the drain contact, and are consequently in equilibrium with the drain $E_F^{\leftarrow} = E_{Fd}$. Since the source and the drain are at the same potential $E_{Fs} - E_{Fd} = qV_{ds} = 0$ V, the right going and left going carriers share a common Fermi level. Notice that we have defined two *quasi*-Fermi levels $E_F^{\rightarrow}$ and $E_F^{\leftarrow}$ and have thus split the carrier distribution into two types that can be in equilibrium amongst themselves, but out of equilibrium with each other. The current is zero at $V_{ds} = 0$ V due to the delicate balance between the left- and right-going current that exactly cancel each other.

This delicate balance is broken when a drain voltage is applied to the transistor.

12.3 Ballistic current-voltage characteristics

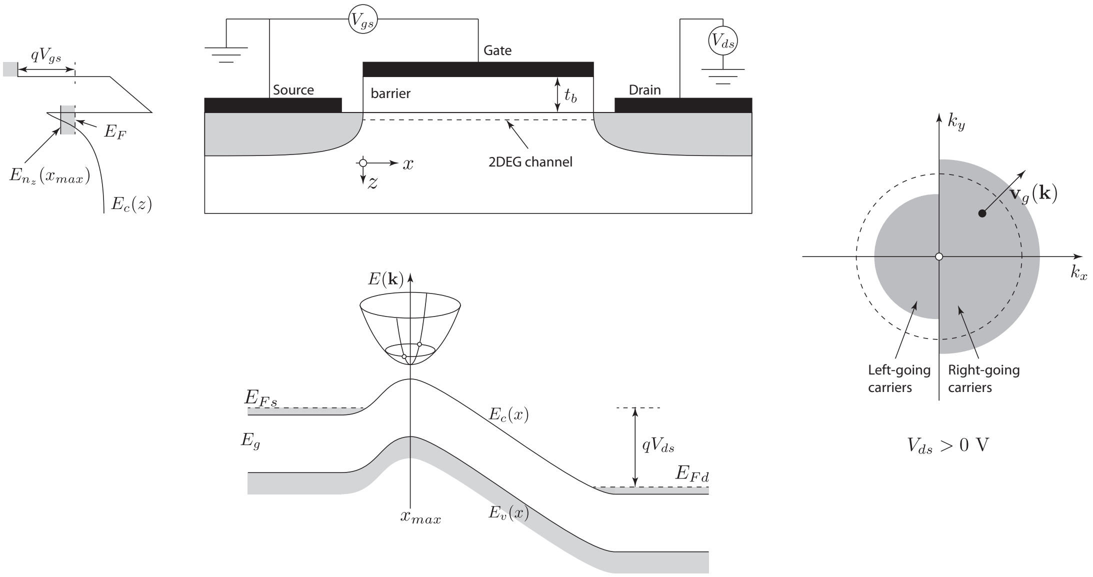


Fig. 12.3: Field effect transistor, energy band diagram, and $\mathbf{k}-$space occupation of states.

When a voltage V_{ds} is applied on the drain, the energy band diagram looks as indicated in Figure 12.1. Now the band edge $E_c(x)$ varies along the channel, with a maximum in the $x - y$ plane occurring at $x = x_{max}$, which is referred to as the 'top-of-the-barrier' (TOB) plane. The ground state of the quantum well $E_{n_z}(x)$ also varies along x depending upon the local vertical electric field, but has the fixed value $E_{n_z}(x_{max})$ at the TOB plane. Interestingly, there is no $x-$oriented electric field at x_{max}. The energy band diagram along the $z-$direction in the TOB plane is also indicated in Figure 12.1. Let's focus on this plane exclusively.

At $V_{ds} = 0$ V, there was a unique E_F at x_{max}, but the quasi-Fermi levels of the right-going carriers and left-going carriers are no longer the same, since $E_{Fs} - E_{Fd} = qV_{ds}$. Due to +ve drain voltage, it has become energetically unfavorable for the drain contact to inject left-going carriers. In the absence of any scattering in the channel, the right-going carriers are still in equilibrium with the source, and the left-going carriers are still in equilibrium with the drain. Thus, the current components now become $J_{2d}^{\rightarrow} = J_0^{2d} F_{1/2}(\eta_s)$ and $J_{2d}^{\leftarrow} = J_0^{2d} F_{1/2}(\eta_d)$. Here $\eta_s = [E_{Fs} - E_{n_z}(x_{max})]/kT$ and $\eta_d = [E_{Fd} - E_{n_z}(x_{max})]/kT = \eta_s - v_d$, where $v_d = qV_{ds}/kT$. The net current of the ballistic transistor is then given by $J_{2d} = J_{2d}^{\rightarrow} - J_{2d}^{\leftarrow}$ as

$$J_{2d} = \frac{qg_s g_v \sqrt{2m^\star}(kT)^{\frac{3}{2}}}{2\pi^2\hbar^2}[F_{\frac{1}{2}}(\eta_s) - F_{\frac{1}{2}}(\eta_s - v_d)] = J_0^{2d}[F_{\frac{1}{2}}(\eta_s) - F_{\frac{1}{2}}(\eta_s - v_d)]. \quad (12.10)$$

The first term is the right-going current carried by the larger gray half-circle in $\mathbf{k}-$apace in Figure 12.1, and the second term is the smaller left-going current carried by the left-going carriers. To evaluate the current, we need to find the dependence of η_s on the gate and drain voltages V_{gs} and V_{ds}.

When $V_{ds} = 0$ V, we found the relation between the unique η and V_{gs} in Eq. 12.6. How do we determine η_s when the carrier distribution looks as in Figure 12.1 with the asymmetric left-and right-going occupation? Here we make the assumption that the net 2DEG density in the TOB plane at $x = x_{max}$ is *completely* controlled by the gate capacitance. This means the net 2DEG density in the TOB plane has not changed from the $V_{ds} = 0$ V case. Experimentally, this is possible when the transistor is electrostatically well designed, with negligible short-channel effects. Let us assume that such design has been achieved.

Then, just like for the current, we split the carrier distribution equation $C_g(V_{gs} - V_T) = C_q V_{th} F_0(\eta)$ from Equation 12.4 into the right-going and left-going carriers as

$$C_g(V_{gs} - V_T) = C_q V_{th} F_0(\eta) \rightarrow C_q V_{th}[\frac{F_0(\eta^{\rightarrow}) + F_0(\eta^{\leftarrow})}{2}]. \quad (12.11)$$

Identifying $\eta^{\rightarrow} = \eta_s$ and $\eta^{\leftarrow} = \eta_s - v_d$ and using $F_0(x) = \ln[1 + \exp(x)]$, we get the relation

$$\ln[(1 + e^{\eta_s})(1 + e^{\eta_s - v_d})] = \frac{2C_g}{C_q}(\frac{V_{gs} - V_T}{V_{th}}) = 2\eta_g = \ln[e^{2\eta_g}], \quad (12.12)$$

which is a quadratic equation in disguise. Solving for η_s yields

$$\eta_s = \ln[\sqrt{(1 + e^{v_d})^2 + 4e^{v_d}(e^{2\eta_g} - 1)} - (1 + e^{v_d})] - \ln[2], \quad (12.13)$$

which reduces to Equation 12.6 for $v_d = 0$. The expression for η_s with $J_{2d}(V_{gs}, V_{ds}) = J_0^{2d}[F_{\frac{1}{2}}(\eta_s) - F_{\frac{1}{2}}(\eta_s - v_d)]$ provides the complete on-state output characteristics of the ballistic FET at any temperature. Note that the expression depends on the values of Fermi-Dirac integrals of order $j = 1/2$. At $V_{ds} = 0$ V, the drain current is zero, as it should be.

Because of the use of Equation 12.11, just as in Equation 12.4, Equation 12.13 works only for the ‘on-state’ of the ballistic transistor. The advantage of this form is that the current can be calculated directly. However, if the off-state characteristics of the ballistic FET are desired, one must find the charge self consistently from Equation 12.7 which read $e^{\frac{qn_s}{C_b V_{th}}}(e^{\frac{qn_s}{C_q V_{th}}} - 1) = e^{\frac{V_{gs} - V_T}{V_{th}}}$ and gave us $n_s(V_{gs})$. Then, the expression to use for the entire ‘on-state’ and ‘off-state’ or sub-threshold behavior of the ballistic FET is simply

$$\eta_s = \ln[\sqrt{(1 + e^{v_d})^2 + 4e^{v_d}(e^{\frac{2n_s(V_{gs})}{n_q}} - 1)} - (1 + e^{v_d})] - \ln[2], \quad (12.14)$$

where we have simply replaced $\eta_g \rightarrow n_s(V_{gs})/n_q$ in Equation 12.13. Based on this general expression, we can evaluate the entire on-state and off-state characteristics of the ballistic FET.

12.4 Examples

The derived expression of the current of the ballistic FET does not depend on the gate length L. This is a consequence of ballistic transport. Figure 12.4 illustrates the entire output characteristics of a ballistic Silicon transistor. The left figure shows the 'transfer' characteristics in log scale, and the middle figure shows the same in linear scale. Note that Equation 12.14 must be used to obtain the on-off switching characteristics exhibited in this figure. Note that the switching is much steeper at a lower temperature, since the subthreshold slope is $\sim 60 \cdot (T/300)$ mV/decade. The right figure shows the drain current per unit width I_d/W as a function of the drain voltage V_{ds}. When V_{ds} is much larger than kT, $v_d >> 1$, and $\eta_s \to \ln[e^{2\eta_g} - 1]$. The current then becomes independent of V_{ds}, i.e., saturates to $J_{2d} \to J_0^{2d} F_{1/2}(\ln[e^{2\eta_g} - 1])$.

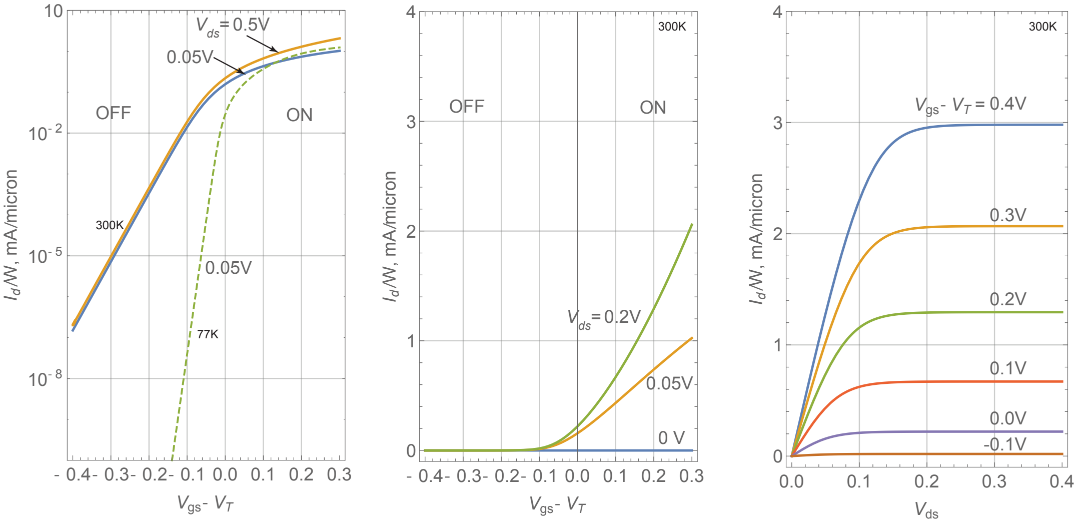


Fig. 12.4: Ballistic Silicon FET. The device dimensions are $t_b = 1$ nm, $\epsilon_b = 10\epsilon_0$, and for Silicon, $m^\star = 0.2m_0$ and $g_v = 2.5$ are used.

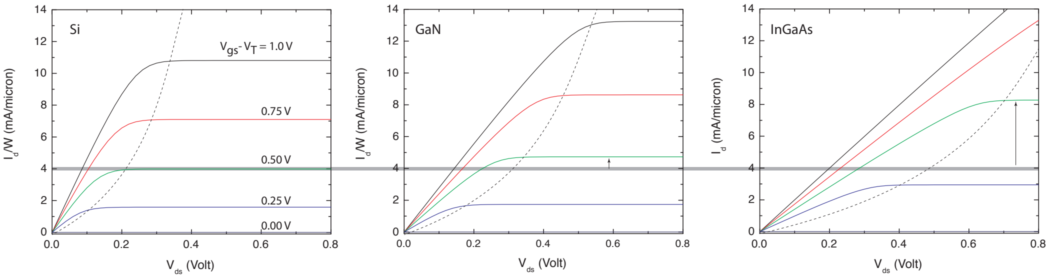


Fig. 12.5: Ballistic FET characteristics at $T = 300$ K for Si, GaN, and $In_{0.53}Ga_{0.47}As$ channels.

The ballistic FET current expression in equation 12.10 is used to plot a few representative cases. The results at room temperature are shown in Figure 12.5. The barrier thickness for all three FETs is chosen to be $t_b = 2$ nm, of a dielectric constant of $\epsilon_b = 10\epsilon_0$. The channel materials chosen are Si, GaN, and $In_{0.53}Ga_{0.47}As$. For Si, an effective valley degeneracy of $g_v = 2.5$, and an effective mass $m^\star \approx 0.2m_0$ is used. For GaN, $g_v = 1$, and

$m^\star \approx 0.2m_0$, and for $\text{In}_{0.53}\text{Ga}_{0.41}\text{As}$ $g_v = 1$, and $m^\star \approx 0.047m_0$ are used. Note that these are representative material parameters, for correlation with experiments, one must make accurate extraction of band parameters from the electronic bandstructures.

The current in Si channels is higher than GaN and $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ channels at low V_{ds}, since it takes advantage of multiple valleys. At high drain bias voltages, the on-current is higher for low effective-mass materials for the same gate overdrive voltage $V_{gs} - V_T$. This boost is due to the higher velocity of carriers due to the low effective mass. For example, at $V_{gs} - V_T = 0.5$ V, the higher saturation currents in GaN and $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ channels are shown by arrows in the Figure. However, it takes higher V_{ds} to attain current saturation.

Due to the ultra thin gate and high gate overdrive voltages, the on-currents predicted are rather high. Experimental highest on-current densities approach ~ 4 mA/micron for nanoscale GaN HEMTs, and lower for Si MOSFETs. The experimental currents are limited by source/drain ohmic contact resistances, and gate leakage. These effects have been neglected in the treatment of the ballistic FET.

However, it is remarkable that even for a ballistic FET with zero source and drain contact resistances and no scattering, the low-V_{ds} regime of the ballistic FET has linear $I_d - V_{ds}$ characteristics and looks like a resistor. One can extract effective on-resistances of the order of ~ 0.05 $\Omega-$mm from the linear regions. The origin of this resistance goes back to the limited number of $|\mathbf{k}\rangle$ states available for transport in the 2DEG channel.

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 13

Through the Barrier: Transmission and Tunneling

Contents

13.1 Tunneling and Transmission 127
13.2 The WKB Method 128
13.3 WKB Method for Semiconductor Transport . 131
13.4 Tunneling transport using the WKB method 132
13.5 Interband 'Zener' Tunneling 133

13.1 Tunneling and Transmission

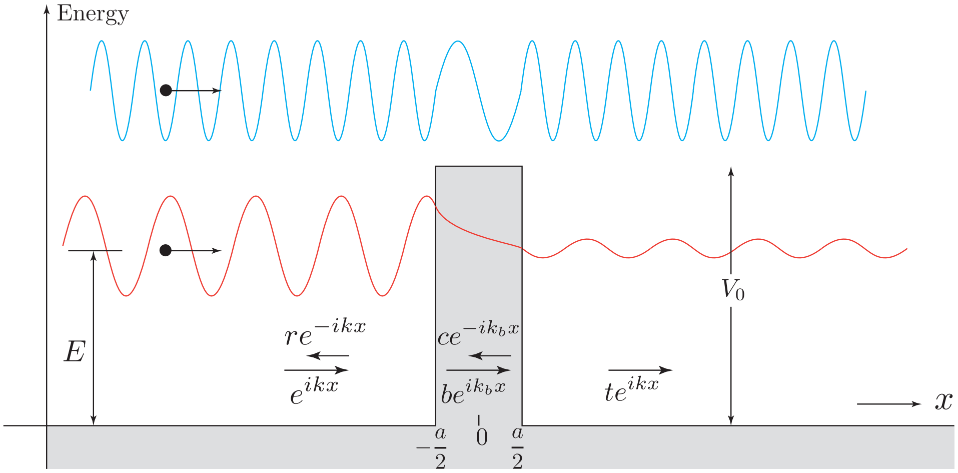


Fig. 13.1: Tunneling and transmission of electrons through a single potential barrier.

Figure 13.1 above shows the transmission of an electron over a barrier, and the quantum mechanical tunneling of an electron through the barrier. For an electron of mass m and kinetic energy E, the boundary conditions at the two barrier interfaces $x = \pm a/2$ give us four equations and four unknowns: the coefficients r, b, c, and t as indicated in the figure. Using boundary conditions that the wavefunction and its derivative are continuous at $x = +a/2$ and at $x = -a/2$ we get 4 equations. Solving these 4 equations yields the 4 unknown coefficients, solving the problem completely.

Since we are interested in the tunneling probability $T(E) = |t|^2$, the above calculation yields

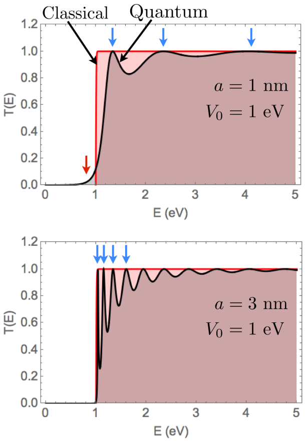


Fig. 13.2: Transmission probability $T = |r|^2$ as a function of the electron energy for a barrier height $V_0 = 1$ eV, and two thicknesses. For electron kinetic energies larger than the barrier height, the transmission probability is given by equation 13.1, and the electron wavevector is real at all x as shown in the wave in blue in Figure 13.1 and the transmission probability is close to one for high energies, with some resonant reflections. For electron kinetic energies lower than the barrier height, the tunneling transmission probability is given by equation 13.2, and as shown in red in Figure 13.1, the amplitude of the function decays exponentially in the barrier and a small amplitude leaks through to the other side.

$$\boxed{T(E) = \frac{1}{1 + \frac{V_0^2}{4E(E-V_0)} \sin^2{(k_b a)}}} \tag{13.1}$$

Note that when the kinetic energy of the electron is lower than the barrier, k_b is imaginary, and $\sin(iy) = i\sinh(y)$. Figure 13.2 shows plots of the tunneling probability as a function of the electron energy E for a fixed barrier height $V_0 = 1$ eV and barrier thicknesses of $a = 10$ nm and $a = 1$ nm. Note how sensitive tunneling is to thickness, and the resonances when $E > V_0$. These values are routinely encountered in semiconductor devices and nanostructures.

$$\boxed{T(E) = \frac{1}{1 + \frac{V_0^2}{4E(V_0-E)} \sinh^2{(k_b a)}}} \approx \frac{16E(V_0-E)}{V_0^2} e^{-2k_b a} \tag{13.2}$$

The boxed expression above is the *exact* tunneling probability for the rectangular barrier, and the approximation will be seen to be equivalent to the WKB approximation, which is the technique that can be applied to potential barriers or arbitrary shapes.

13.2 The WKB Method

The WKB method is a recipe for solving the time-independent Schrodinger equation for arbitrary potentials. There is typically a misconception that this method is a crude approximation. Far from it. Consider the time-independent Schrodinger equation in 1D:

$$-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2}\psi(x) + V(x)\psi(x) = E\psi(x) \tag{13.3}$$

rewritten in the form

$$\psi''(x) = Q(x)\psi(x), \tag{13.4}$$

with

$$Q(x) = \frac{2m_e}{\hbar^2}(V(x) - E). \tag{13.5}$$

This is in general analytically unsolvable, except for the few cases considered in this chapter such as the harmonic oscillator, or the hydrogen atom. George Green (the same person who is credited with Green's functions) had a brilliant insight to crack into an asymptotic approach to a solution. He suggested that we try a solution of the form

$$\psi(x) = e^{S(x)} \tag{13.6}$$

Substituting, we obtain $(S'' + (S')^2)e^S = Qe^S$ which leads to

$$S'' + (S')^2 = Q \tag{13.7}$$

Note that this is *exact*. But unfortunately we have just converted the already unsolvable Schrodinger equation into another insolvable equation: the Riccati equation. It is also a non-linear equation!

Green made the crucial observation that if $S(x) = ax^b$ where $b < 0$, then $S'' << (S')^2$ for small x. This makes the asymptotic form of Eq. 13.7

$$(S')^2 \sim Q, \tag{13.8}$$

which leads to

$$S' \sim \pm\sqrt{Q}, \tag{13.9}$$

and

$$S(x) \sim \pm \int_a^x du\sqrt{Q(u)}. \tag{13.10}$$

Here a is any chosen constant. Now lets write

$$S(x) = \underbrace{\pm \int_a^x du\sqrt{Q(u)}}_{S_0(x)} + \underbrace{C(x)}_{S_1(x)+S_2(x)+S_3(x)+\ldots} \tag{13.11}$$

Now we have gone back from an asymptotic form to an exact form here. Note that the leading functional form is defined $S_0(x) = \pm\int_a^x du\sqrt{Q(u)}$. Consequently, we have the relations

$$S_0^{'}(x) = \pm\sqrt{Q(x)} \tag{13.12}$$

$$S_0^{''}(x) = \pm\frac{Q'(x)}{2\sqrt{Q(x)}} \tag{13.13}$$

Now we substitute $S = S_0 + C$ into 13.7 to get

$$S_0^{''} + C'' + (S_0^{'} + C')^2 = Q. \tag{13.14}$$

Note that this is back to being *exact*. Writing it out, we get

$$S_0^{''} + C'' + (S_0^{'})^2 + (C')^2 + 2S_0^{'}C' = Q, \tag{13.15}$$

and substituting all terms related to $S_0(x)$ and its derivatives, we get

$$\pm\frac{Q'(x)}{2\sqrt{Q(x)}} + C'' + Q + (C')^2 \pm 2\sqrt{Q(x)}C' = Q. \tag{13.16}$$

This is still *exact*. Note the crucial cancellation of Q. The result is a differential equation, but *now for the correction function* $C(x)$:

$$\pm\frac{Q'(x)}{2\sqrt{Q(x)}} + C'' + (C')^2 \pm 2\sqrt{Q(x)}C' = 0 \tag{13.17}$$

But this equation is not very different from the Riccati equation either! Now we step back to asymptotics again: since $S_0(x)$ is the leading order or dominant function, we have the relations $C(x) << \pm\int_a^x dt\sqrt{Q(t)}$, and consequently

$$C'(x) << \pm\sqrt{Q(x)}, \tag{13.18}$$

$$C''(x) << \pm\frac{Q'(x)}{2\sqrt{Q(x)}}, \tag{13.19}$$

As a result, we may throw away the C'' and $(C')^2$ terms to obtain:

$$C'(x) \sim -\frac{Q'(x)}{4Q(x)}, \tag{13.20}$$

which yields

$$C(x) \sim -\frac{1}{4}\ln Q(x) = \underbrace{-\frac{1}{4}\ln Q(x)}_{S_1(x)} + S_2(x) + S_3(x) + ..., \tag{13.21}$$

Thus, the solution to the Schrodinger equation is

$$\psi(x) = e^{S(x)} = e^{S_0(x)+S_1(x)+S_2(x)...} = e^{S_0(x)}e^{S_1(x)}e^{S_2(x)+...} \tag{13.22}$$

Substituting the values of S_0 and S_1, we obtain

$$\psi(x) = e^{S(x)} = e^{\pm\int_a^x du\sqrt{Q(u)}}e^{-\frac{1}{4}\ln Q(x)}e^{S_2(x)+...} \tag{13.23}$$

which is rewritten in the form

$$\psi(x) = e^{S(x)} = \frac{1}{(Q)^{\frac{1}{4}}}e^{\pm\int_a^x du\sqrt{Q(u)}}e^{S_2(x)+...} \tag{13.24}$$

This process may be repeated again and again. Each step will generate new functional forms $S_2(x), S_3(x)$.... Green's method extracts out the "non-Frobenius" and "non-Fuchs" terms, and thereafter the series becomes rational powers of x: $D(x) = S_2(x) + S_3(x) + ... = \sum_n a_n(\sqrt{x})^n$. But this is rarely necessary in our work. For most cases, it turns out that $D(x) << 1$, and the 'WKB' approximation to use looks like

$$\boxed{\psi(x) \approx \frac{K}{Q(x)^{\frac{1}{4}}}e^{\pm\int_a^x du\sqrt{Q(u)}}.} \tag{13.25}$$

where K is a constant.

One can *mechanize* the extraction of the successive functions $S_n(x)$ by the following *perturbative* approach. We recast the Schrodinger equation as a function of a parameter ϵ of the form $\epsilon^2\psi'' = Q\psi$, and assume a solution

$$\psi(x) = e^S = e^{\frac{1}{\epsilon}\sum_{n=0}^{\infty} S_n\epsilon^n} = e^{\frac{1}{\epsilon}S_0+S_1+S_2\epsilon+S_3\epsilon^3+...} \tag{13.26}$$

Substituting this form into the Schrodinger equation $\epsilon^2\psi'' = Q\psi$ leads to the set of equations defined by

$$(S_0' + S_1'\epsilon + S_2'\epsilon^2 + S_3'\epsilon^3)^2 + (S_0''\epsilon + S_1''\epsilon^2 + S_2''\epsilon^3 + ...) = Q, \tag{13.27}$$

from where we equate like powers of ϵ on both sides of the equation to obtain

$$(S_0')^2 = Q \quad \text{for } \epsilon^0 \tag{13.28}$$

$$2S_0'S_1' + S_0'' = 0 \quad \text{for } \epsilon^1 \tag{13.29}$$

$$2S_0'S_2' + S_1'' + (S_1')^2 = 0 \quad \text{for } \epsilon^2 \tag{13.30}$$

$$2S_0'S_n' + S_{n-1}'' + \sum_{j=1}^{n-1} S_j'S_{n-j}' = 0 \quad \text{for } \epsilon^n \tag{13.31}$$

So what is this ϵ? One way to look at it is to 'relate' it to $\hbar$, which is a small number. But ϵ is a variable in the above formulation, which renders the Schrodinger equation into a polynomial series, enabling the WKB evaluation.

13.3 WKB Method for Semiconductor Transport

Now consider the effective mass equation for an electron in the conduction band of a semiconductor where the band edge $E_c(x)$ varies smoothly in space. Then, the effective mass equation for the electron is

$$[E_n(-i\nabla)]C(r) = [E - E_c(x)]C(x) \implies \frac{d^2}{dx^2}C(x) = \underbrace{-\frac{2m_c^\star}{\hbar^2}[E - E_c(x)]}_{Q(x)} C(x), \quad (13.32)$$

where $C(x)$ is the envelope function, and the full wavefunction of the wavepacket representing the electron is $\psi(x) \approx C(x)u_k(x)$. This is clearly in the form of the WKB formulation; the solution is

$$\boxed{C(x) \approx \frac{K}{Q(x)^{\frac{1}{4}}} e^{\pm \int_a^x du\sqrt{Q(u)}}.} \quad (13.33)$$

where K is a constant. Now if the energy of the wavepacket is larger than the conduction band edge $E > E_c(x)$, $Q(x) < 0$ and $\sqrt{Q(x)} = ik(x)$, where $k(x) = \sqrt{\frac{2m_c^\star}{\hbar^2}(E - E_c(x))}$ is a spatially varying wavevector. Then, the envelope function varies in space as

$$\boxed{C(x) \approx \frac{K'}{\sqrt{k(x)}} e^{\pm i \int_a^x du k(u)},} \quad (13.34)$$

which looks like a plane wave whose wavelength is smoothly varying in space, and the amplitude is decreasing as the wavevector increases. This situation is shown in Figure 13.3.

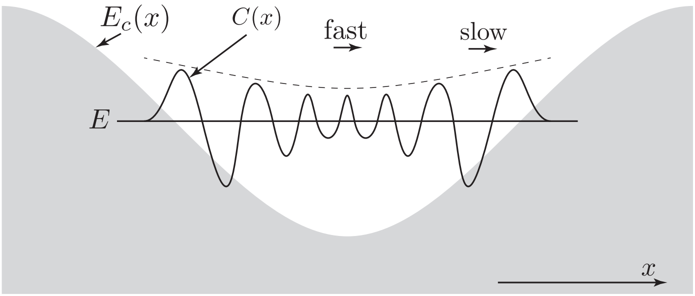


Fig. 13.3: Electron transport and approximate wavefunction in smoothly varying potentials.

The wavevector $k(x) = \sqrt{\frac{2m_c^\star}{\hbar^2}(E - E_c(x))}$ is large for x where the net kinetic energy $E - E_c(x)$ is large; the electron is moving fast at those points. Near the classical turning points where $E - E_c(x) \to 0$, $k(x)$ decreases, the wavelength stretches out, and the electron slows down. The classical analogy of a ball rolling up and down the valley should be clear. Just as in the classical situation, the *probability* of finding the particle is low at x locations where it is moving fast. This feature is captured in the quantum mechanical wavefunction, since the probability density goes as $|C(x)|^2 \propto \frac{1}{k(x)}$ as indicated by the dashed line.

The dynamics of a quantum particle represented by the wavefunction $\psi(x,t)$ is governed by the time-dependent Schrodinger equation

$$[-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2}+V(x)]\psi(x,t)=i\hbar\frac{\partial}{\partial t}\psi(x,t) \implies j=\frac{\psi^\star\hat{p}\psi-\psi\hat{p}\psi^\star}{2m_e} \tag{13.35}$$

.

where the quantum mechanical probability current density j is obtained from the wavefunction. Note that the electron mass is in the denominator, and ψ is the complete wavefunction. Now we recognize the effective mass equation $E_c(-i\nabla)=E_c(x)-\frac{\hbar^2}{2m_c^\star}\frac{d^2}{dx^2}$ which makes the effective mass equation mathematically identical to the free-electron Schrodinger equation:

$$[-\frac{\hbar^2}{2m_c^\star}\frac{d^2}{dx^2}+E_c(x)]C(x,t)=i\hbar\frac{\partial}{\partial t}C(x,t) \implies j=-\frac{i\hbar}{2m_c^\star}[C^\star\frac{\partial C}{\partial x}-C\frac{\partial C^\star}{\partial x}]. \tag{13.36}$$

and the quantum mechanical probability current density j is thus obtained directly from the effective mass parameters, by using the envelope function $C(x)$ and the effective mass $m_c^\star$ instead of the free electron mass.

Now since the charge current carried by an effective mass state is

$$J=qg_sg_v\sum_k j_k=g_sg_v\sum_k v_g(k)|C(x)|^2, \tag{13.37}$$

we get from equation 13.34 for a purely right-going WKB state $C(x)=\frac{Ke^{i\int_a^x k(u)du}}{\sqrt{k(x)}}$ in a smoothly varying conduction band profile $E_c(x)$. Because near the band edge the bandstructure is $E_c(k)=E_c(x)+\frac{\hbar^2k(x)^2}{2m_c^\star}$, the group velocity of the effective mass state is $v_g(k)=\frac{1}{\hbar}\frac{\partial E_c(k)}{\partial k}=\frac{\hbar k(x)}{m_c^\star}$ The charge current density is then found to be

$$J=qg_sg_v\underbrace{\frac{\hbar k(x)}{m_c^\star}}_{v(x)}\underbrace{|C(x)|^2}_{n(x)}=qg_sg_v\frac{\hbar k(x)}{m_c^\star}\frac{|K|^2}{k(x)}=qg_sg_v\frac{\hbar|K|^2}{m_c^\star}, \tag{13.38}$$

showing how the net current is conserved in transport across a smoothly varying potential profile. This is shown in Figure 13.4 for an effective mass electron wavepacket in the conduction band of a semiconductor. The envelope carrier density $n(x)=C^\star(x)C(x)=\frac{|K|^2}{k(x)}$ decreases when the group velocity $\frac{\hbar k(x)}{m_c^\star}$ increases, keeping their product $J(x)\sim n(x)v(x)$ constant.

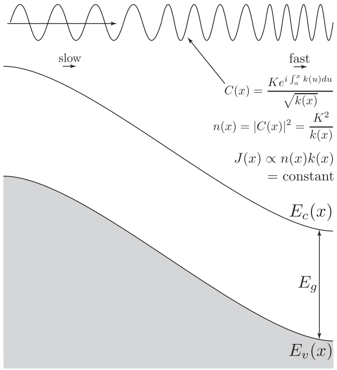

slow
fast
$C(x)=\frac{Ke^{i\int_a^x k(u)du}}{\sqrt{k(x)}}$
$n(x)=|C(x)|^2=\frac{K^2}{k(x)}$
$J(x)\propto n(x)k(x)$
$=$ constant
$E_c(x)$
E_g
$E_v(x)$

Fig. 13.4: Transport of an effective mass electron wavepacket in the conduction band in a smoothly varying potential. The group velocity $v(x)\sim k(x)$ increases as the kinetic energy increases, but the carrier density $n(x)=|C(x)|^2\sim 1/k(x)$ decreases, keeping the net current $J(x)\sim n(x)v(x)$ constant.

13.4 Tunneling transport using the WKB method

The electron may encounter potential barriers whose height is larger than the kinetic energy, i.e. $E<V(x)$ as indicated in Figure 13.5. In such a situation, the term $Q(x)=\kappa(x)=-\frac{2m_c^\star}{\hbar^2}(E-E_c(x))>0$, and the WKB solution of Equation 13.33 is

$$\boxed{C(x)\approx\frac{K}{\sqrt{\kappa(x)}}e^{\pm\int_a^x\kappa(u)du}.} \tag{13.39}$$

If the classical turning points are $x=x_1$ and $x=x_2$ as indicated in Figure 13.5, we can write the ratio of the coefficients approximately as

$$\frac{C(x_2)}{C(x_1)}\approx\frac{e^{-\int_a^{x_2}\kappa(u)du}}{e^{-\int_a^{x_1}\kappa(u)du}}\approx e^{i\phi}e^{-\int_{x_1}^{x_2}\kappa(u)du}, \tag{13.40}$$

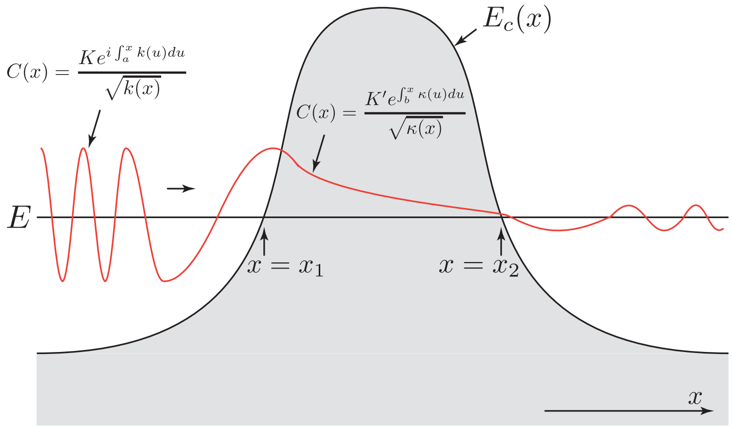

$E_c(x)$
$C(x) = \frac{Ke^{i\int_a^x k(u)du}}{\sqrt{k(x)}}$
$C(x) = \frac{K'e^{\int_b^x \kappa(u)du}}{\sqrt{\kappa(x)}}$
E
$x = x_1$
$x = x_2$
x

Fig. 13.5: WKB tunneling through a potential barrier.

where ϕ is a possible phase factor accumulated in the oscillatory part. Because the tunneling probability is proportional to the square of the wavefunction, we obtain

$$|\frac{C(x_2)}{C(x_1)}|^2 \approx e^{-2\int_{x_1}^{x_2} \kappa(u)du} \implies \boxed{T_{wkb} \approx e^{-2\int_{x_1}^{x_2} dx\sqrt{\frac{2m_c^\star}{\hbar^2}[E_c(x)-E]}}}. \qquad (13.41)$$

This approximate equation holds for tunneling within a single band. It is not exact, but provides a reasonable estimate of the tunneling probability. It may be compared with the exact tunneling probability obtained in the rectangular barrier in Equation 13.2. Using the physical constants, we obtain $T_{wkb} \approx e^{-(\frac{t_b}{0.1 \text{ nm}})\sqrt{(\frac{m_c^\star}{m_e})\cdot(\frac{V_0}{1 \text{ eV}})}}$. For example, the tunneling probability for a barrier $t_b = 1$ nm thick and uniform barrier height $V_0 \sim 1$ eV, with a conduction band edge effective mass $m_c^\star \sim 0.09m_e$ is $T_{wkb} \approx e^{-3} \sim 1/20$, rather high number because of the thin barrier and the light effective mass. The tunneling probability is extremely sensitive to the thickness, barrier height and the effective mass. This sensitivity is also a reason the tunneling currents can be used to measure these parameters.

13.5 Interband 'Zener' Tunneling

At high electric fields In semiconductors, tunneling of electrons can occur from the valence band to the conduction band. Such interband tunneling may only be loosely modeled by the above effective mass formalism, because the effective mass approximation works best for electron transport within the same band. Let us first develop a very approximate method to estimate the interband tunneling currents, and then refine it in successive steps.

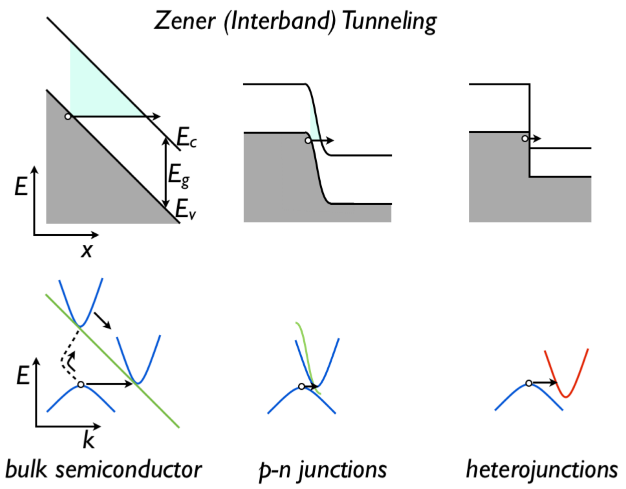
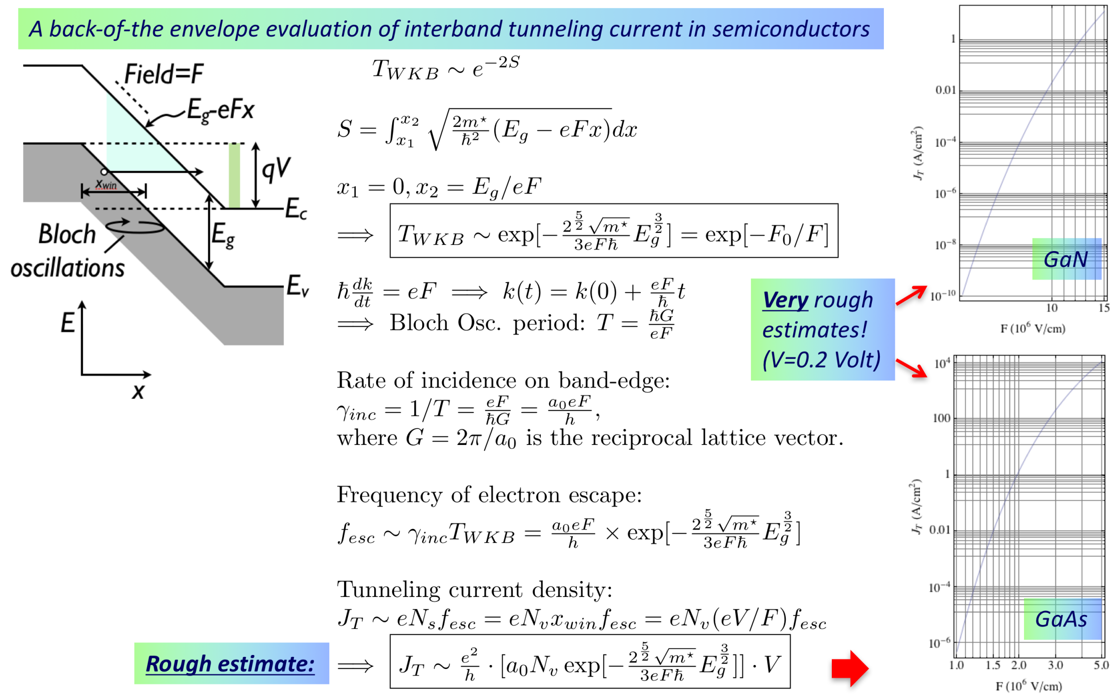


Fig. 13.6: Interband Zener tunneling in bulk semiconductors, p-n junctions, and heterojunctions.



A back-of-the envelope evaluation of interband tunneling current in semiconductors

Field=F
E_g-eFx
qV
x_{win}
E_c
E_g
E_v
Bloch
oscillations
E
x

$$T_{WKB} \sim e^{-2S}$$

$$S = \int_{x_1}^{x_2} \sqrt{\frac{2m^\star}{\hbar^2}(E_g - eFx)}dx$$

$$x_1 = 0, x_2 = E_g/eF$$

$$\implies \boxed{T_{WKB} \sim \exp[-\frac{2^{\frac{5}{2}}\sqrt{m^\star}}{3eF\hbar}E_g^{\frac{3}{2}}] = \exp[-F_0/F]}$$

$$\hbar\frac{dk}{dt} = eF \implies k(t) = k(0) + \frac{eF}{\hbar}t$$
$$\implies \text{Bloch Osc. period: } T = \frac{\hbar G}{eF}$$

Rate of incidence on band-edge:
$\gamma_{inc} = 1/T = \frac{eF}{\hbar G} = \frac{a_0 eF}{h}$,
where $G = 2\pi/a_0$ is the reciprocal lattice vector.

Frequency of electron escape:
$f_{esc} \sim \gamma_{inc} T_{WKB} = \frac{a_0 eF}{h} \times \exp[-\frac{2^{\frac{5}{2}}\sqrt{m^\star}}{3eF\hbar}E_g^{\frac{3}{2}}]$

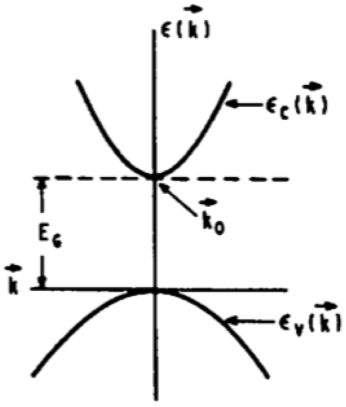
Tunneling current density:
$J_T \sim eN_s f_{esc} = eN_v x_{win} f_{esc} = eN_v(eV/F)f_{esc}$

Rough estimate: $\implies \boxed{J_T \sim \frac{e^2}{h} \cdot [a_0 N_v \exp[-\frac{2^{\frac{5}{2}}\sqrt{m^\star}}{3eF\hbar}E_g^{\frac{3}{2}}]] \cdot V}$

***Very** rough estimates! (V=0.2 Volt)*

Fig. 13.7: A simple model for calculating the interband tunneling current density in semiconductors.

Conservation of lateral momentum:
(since there are no forces in that direction)
→ Holds for direct-gap semiconductors.
→ Phonons necessary for indirect gap Scs.



Duke (Tunneling in Solids)

Tunneling probability in 3D semiconductors:

$$\boxed{T(\mathbf{k}_\parallel) = T_0 \exp[-c_0 \frac{\sqrt{m^\star}}{\hbar e F} E_g^{3/2}] \exp[-\frac{E_\parallel}{\bar{E}_\parallel}]},$$

where T_0 and c_0 are of the order of 1.

$$E_\parallel = \frac{\hbar^2(\mathbf{k}_\parallel - \mathbf{k}_{0,\parallel})^2}{2m^\star}$$

Electrons in states with $\mathbf{k}_\parallel \neq 0$ see an effectively larger gap, and their tunneling probability is hence exponentially damped with energy. Expressions for $\bar{E}_\parallel$ depend on the chosen bandstructure model.

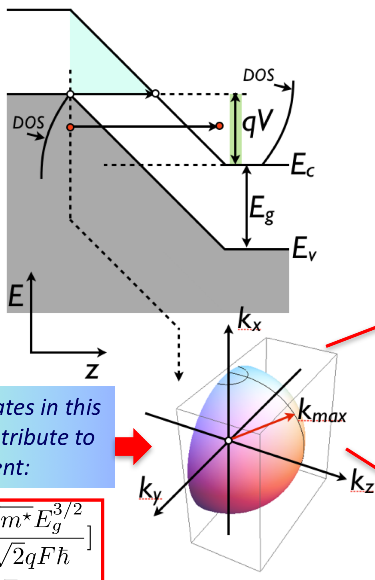
For a 2-band $\mathbf{k}\cdot\mathbf{p}$ model: $\begin{pmatrix} E_g + \frac{\hbar^2 k^2}{2m_0} & \frac{\hbar \mathbf{k}\cdot\mathbf{p}}{m_0} \\ \frac{\hbar \mathbf{k}\cdot\mathbf{p}}{m_0} & \frac{\hbar^2 k^2}{2m_0} \end{pmatrix}$

$$\boxed{T_0 = (\frac{\pi}{3})^2, c_0 = \frac{\pi}{2^{3/2}}, m^\star = \frac{m_0^2 E_g}{2p_{cv}^2}, \bar{E}_\parallel = \frac{2^{1/2}\hbar e F}{2\pi\sqrt{m^\star}\sqrt{E_g}}},$$

As before, the momentum matrix element of most semiconductors satisfy the relation $\boxed{2p_{cv}^2/m_0 \approx 20 \text{ eV}}$.

Fig. 13.8: Interband tunneling probabilities.

Procedure for calculating tunneling current:



A part of the states in this hemisphere contribute to the tunnel current:

$$\boxed{\begin{aligned} T_0 &= \exp[-\frac{\pi\sqrt{m^\star}E_g^{3/2}}{2\sqrt{2}qF\hbar}] \\ \bar{E}_\parallel &= \frac{\sqrt{2}qF\hbar}{2\pi\sqrt{m^\star}\sqrt{E_g}} \end{aligned}}$$

Reverse-bias Tunnel current:

Energy + lateral momentum conservation:

$$E_v^p - \frac{\hbar^2}{2m_v^\star}(k_\parallel^2 + k_z^2) = E_c^n + \frac{\hbar^2}{2m_c^\star}(k_\parallel^2 + (k_z')^2)$$

$$\boxed{\implies \frac{\hbar^2}{2\mu}k_\parallel^2 + \frac{\hbar^2}{2m_v^\star}k_z^2 + \frac{\hbar^2}{2m_c^\star}(k_z')^2 = E_v^p - E_c^n = qV},$$

where $\mu = m_c^\star m_v^\star/(m_c^\star + m_v^\star)$.

If we assume $m_v^\star = m_c^\star = m^\star = 2\mu$, we get the modified energy conservation relation:

$$\boxed{2k_\parallel^2 + k_z^2 = \frac{2m^\star}{\hbar^2}qV - (k_z')^2}.$$

Denote $\frac{2m^\star}{\hbar^2}qV = k_{max}^2$. Now $(k_z')^2 \geq 0$, and

$$\boxed{2k_\parallel^2 + k_z^2 \leq k_{max}^2}$$

is the surface bounding the volume of states in the $k-$space that contribute to tunneling current.

$$J = \frac{2e}{V}\sum_k v_g^z(\mathbf{k})T(\mathbf{k})[f_L(\mathbf{k}) - f_R(\mathbf{k})].$$

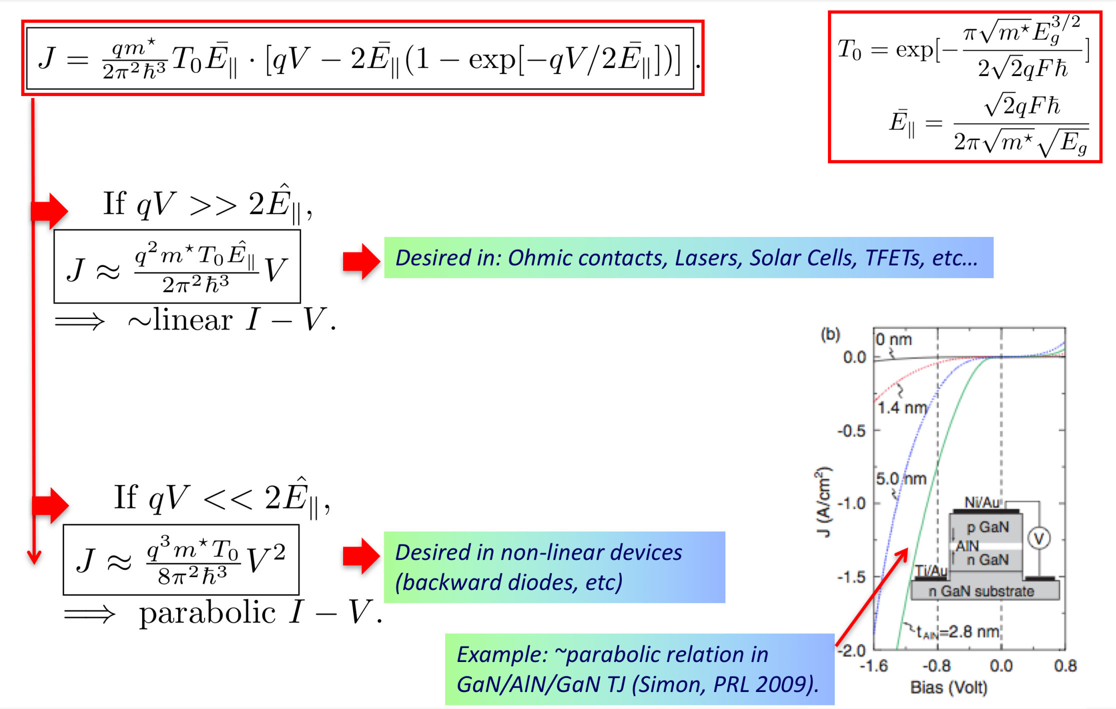
Spherical coordinates: $k_z = k\cos\theta$, $k_\parallel = k\sin\theta$. Energy/momentum conservation volume Ω: $k^2 \leq k_{max}^2/(1+\sin^2\theta)$. Then,

$$J = \frac{2e}{(2\pi)^3}\int_\Omega d^3k v_g^z(\mathbf{k})T(\mathbf{k})[f_L(\mathbf{k}) - f_R(\mathbf{k})],$$

and we get:

$$\boxed{J = \frac{qm^\star}{2\pi^2\hbar^3}T_0\bar{E}_\parallel \cdot [qV - 2\bar{E}_\parallel(1 - \exp[-qV/2\bar{E}_\parallel])]}.$$

Fig. 13.9: Calculation of interband tunneling current with lateral momentum conservation.



$$\boxed{J = \frac{qm^\star}{2\pi^2\hbar^3} T_0 \bar{E}_\parallel \cdot [qV - 2\bar{E}_\parallel(1-\exp[-qV/2\bar{E}_\parallel])]}.$$

$$T_0 = \exp[-\frac{\pi\sqrt{m^\star}E_g^{3/2}}{2\sqrt{2}qF\hbar}]$$

$$\bar{E}_\parallel = \frac{\sqrt{2}qF\hbar}{2\pi\sqrt{m^\star}\sqrt{E_g}}$$

If $qV >> 2\hat{E}_\parallel$,

$$\boxed{J \approx \frac{q^2 m^\star T_0 \hat{E}_\parallel}{2\pi^2\hbar^3}V}$$

$\implies$ ~linear $I-V$.

Desired in: Ohmic contacts, Lasers, Solar Cells, TFETs, etc...

If $qV << 2\hat{E}_\parallel$,

$$\boxed{J \approx \frac{q^3 m^\star T_0}{8\pi^2\hbar^3}V^2}$$

$\implies$ parabolic $I-V$.

Desired in non-linear devices (backward diodes, etc)

Example: ~parabolic relation in GaN/AlN/GaN TJ (Simon, PRL 2009).

Fig. 13.10: Interband tunneling current densities at low and high reverse bias voltages.

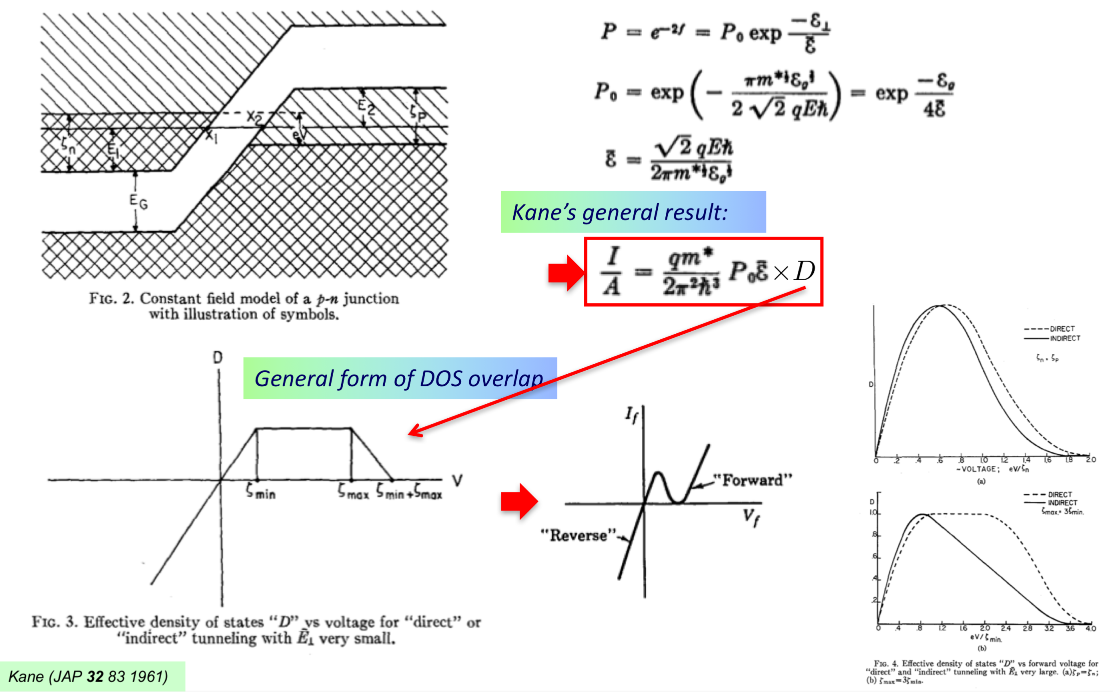


Fig. 2. Constant field model of a *p-n* junction with illustration of symbols.

$$P = e^{-2f} = P_0 \exp\frac{-\mathcal{E}_\perp}{\bar{\mathcal{E}}}$$

$$P_0 = \exp\left(-\frac{\pi m^{*\frac{1}{2}}\mathcal{E}_g^{\frac{3}{2}}}{2\sqrt{2}\,qE\hbar}\right) = \exp\frac{-\mathcal{E}_g}{4\bar{\mathcal{E}}}$$

$$\bar{\mathcal{E}} = \frac{\sqrt{2}\,qE\hbar}{2\pi m^{*\frac{1}{2}}\mathcal{E}_g^{\frac{1}{2}}}$$

Kane's general result:

$$\boxed{\frac{I}{A} = \frac{qm^*}{2\pi^2\hbar^3}P_0\bar{\mathcal{E}} \times D}$$

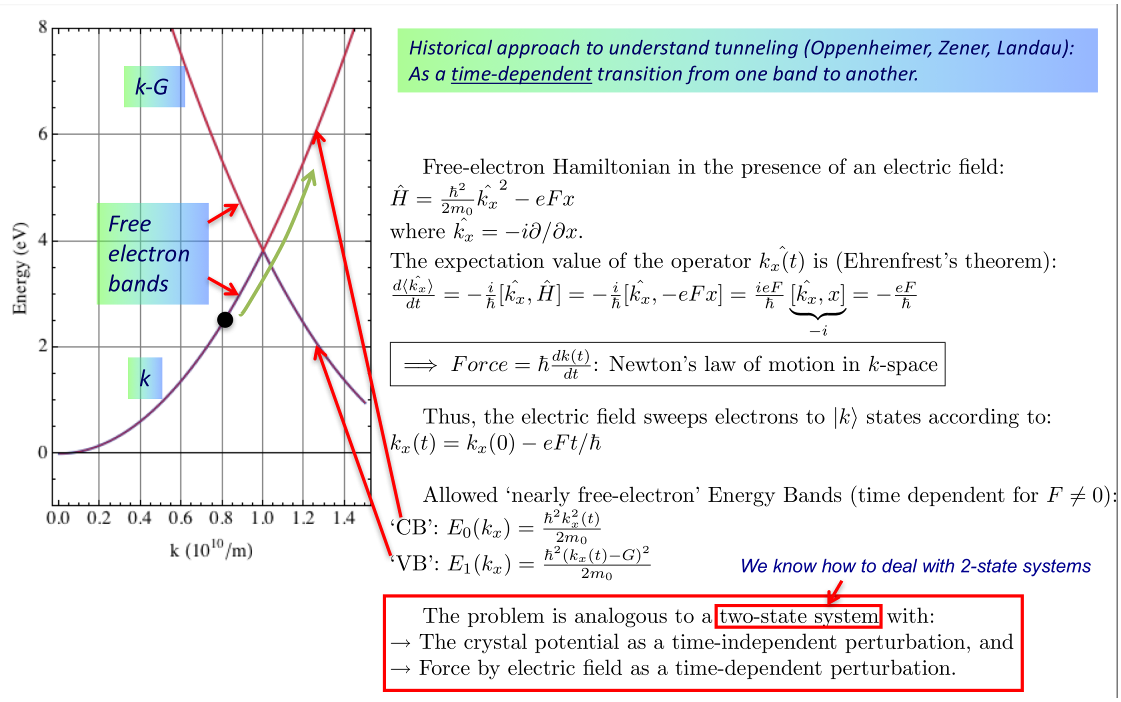
General form of DOS overlap

Fig. 3. Effective density of states "D" vs voltage for "direct" or "indirect" tunneling with $\bar{E}_\perp$ very small.

Fig. 4. Effective density of states "D" vs forward voltage for "direct" and "indirect" tunneling with $\bar{E}_\perp$ very large. (a) $\zeta_p = \zeta_n$; (b) $\zeta_{max} = 3\zeta_{min}$.

Kane (JAP **32** 83 1961)

Fig. 13.11: Negative differential resistance in interband tunneling in degenerately doped Esaki diodes.



Historical approach to understand tunneling (Oppenheimer, Zener, Landau): As a time-dependent transition from one band to another.

Free-electron Hamiltonian in the presence of an electric field:
$\hat{H} = \frac{\hbar^2}{2m_0}\hat{k_x}^2 - eFx$
where $\hat{k_x} = -i\partial/\partial x$.
The expectation value of the operator $\hat{k_x(t)}$ is (Ehrenfrest’s theorem):
$\frac{d\langle\hat{k_x}\rangle}{dt} = -\frac{i}{\hbar}[\hat{k_x}, \hat{H}] = -\frac{i}{\hbar}[\hat{k_x}, -eFx] = \frac{ieF}{\hbar}\underbrace{[\hat{k_x}, x]}_{-i} = -\frac{eF}{\hbar}$

$\boxed{\implies Force = \hbar\frac{dk(t)}{dt}\text{: Newton's law of motion in } k\text{-space}}$

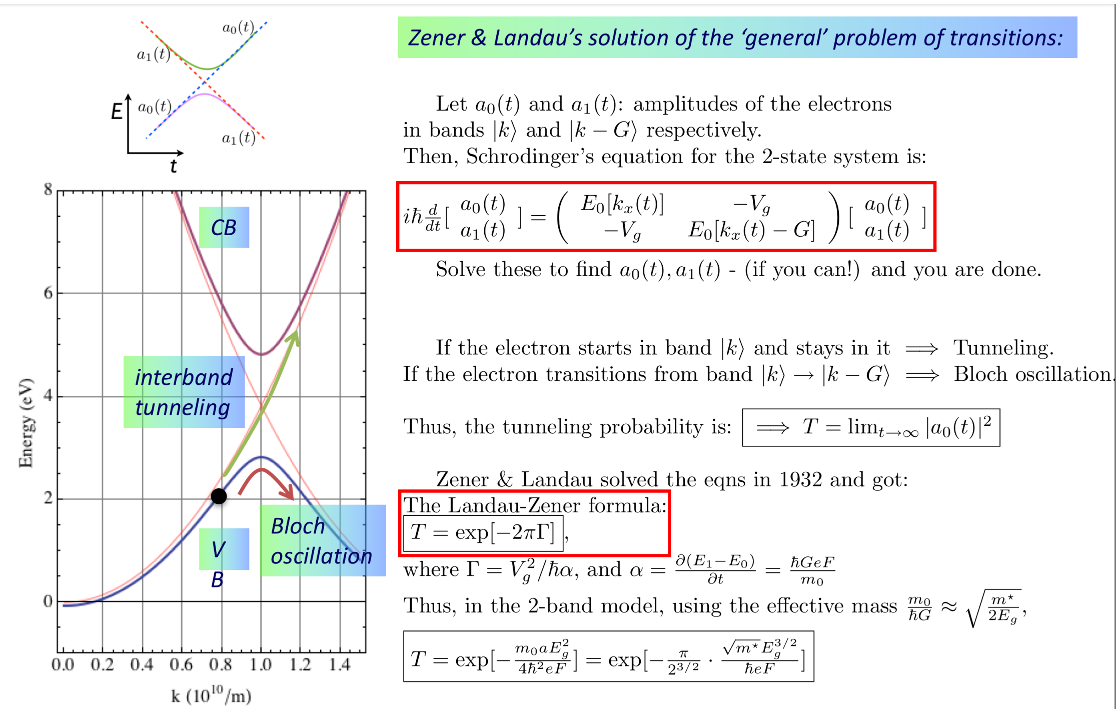
Thus, the electric field sweeps electrons to $|k\rangle$ states according to:
$k_x(t) = k_x(0) - eFt/\hbar$

Allowed ‘nearly free-electron’ Energy Bands (time dependent for $F \neq 0$):
‘CB’: $E_0(k_x) = \frac{\hbar^2 k_x^2(t)}{2m_0}$
‘VB’: $E_1(k_x) = \frac{\hbar^2 (k_x(t)-G)^2}{2m_0}$

We know how to deal with 2-state systems

The problem is analogous to a two-state system with:
→ The crystal potential as a time-independent perturbation, and
→ Force by electric field as a time-dependent perturbation.

Fig. 13.12: The Landau-Zener approach to interband tunneling as a time-dependent perturbation problem.



Zener & Landau’s solution of the ‘general’ problem of transitions:

Let $a_0(t)$ and $a_1(t)$: amplitudes of the electrons in bands $|k\rangle$ and $|k-G\rangle$ respectively.
Then, Schrodinger’s equation for the 2-state system is:

$$i\hbar\frac{d}{dt}\left[\begin{array}{c} a_0(t) \\ a_1(t) \end{array}\right] = \begin{pmatrix} E_0[k_x(t)] & -V_g \\ -V_g & E_0[k_x(t)-G] \end{pmatrix}\left[\begin{array}{c} a_0(t) \\ a_1(t) \end{array}\right]$$

Solve these to find $a_0(t), a_1(t)$ - (if you can!) and you are done.

If the electron starts in band $|k\rangle$ and stays in it $\implies$ Tunneling.
If the electron transitions from band $|k\rangle \to |k-G\rangle$ $\implies$ Bloch oscillation.

Thus, the tunneling probability is: $\boxed{\implies T = \lim_{t\to\infty}|a_0(t)|^2}$

Zener & Landau solved the eqns in 1932 and got:
The Landau-Zener formula:
$\boxed{T = \exp[-2\pi\Gamma]}$,
where $\Gamma = V_g^2/\hbar\alpha$, and $\alpha = \frac{\partial(E_1-E_0)}{\partial t} = \frac{\hbar GeF}{m_0}$
Thus, in the 2-band model, using the effective mass $\frac{m_0}{\hbar G} \approx \sqrt{\frac{m^\star}{2E_g}}$,

$$\boxed{T = \exp[-\frac{m_0 aE_g^2}{4\hbar^2 eF}] = \exp[-\frac{\pi}{2^{3/2}}\cdot\frac{\sqrt{m^\star}E_g^{3/2}}{\hbar eF}]}$$

Fig. 13.13: The Landau-Zener approach to interband tunneling as a time-dependent perturbation problem.

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 14

Fermi's Golden Rule

Contents

14.1 Introduction 139
14.2 Fermi's Golden Rule . . . 139
14.3 Perturbations oscillating in time 142
14.4 Transitions to a continuum of states . . . 143
14.5 Higher order transitions: Dyson series and Diagrams 144
14.6 Fate of the initial state: Self Energy 145

14.1 Introduction

In this chapter, we derive a very useful result for estimating transition rates between quantum states due to time-dependent perturbation. The results will be used heavily in subsequent chapters to understand the optical and electronic transport properties of semiconductors.

14.2 Fermi's Golden Rule

Consider an unperturbed quantum system in state $|\Psi_{t_0}\rangle$ at time $t = t_0$. It evolves to the state $|\Psi_t\rangle$ at a future instant t. The time evolution of the state vector is governed by the unperturbed Hamiltonian H_0 according to the time-dependent Schrodinger equation

$$i\hbar\frac{\partial}{\partial t}|\Psi_t\rangle = H_0|\Psi_t\rangle. \tag{14.1}$$

If the system was in an eigenstate $|\Psi_{t_0}\rangle = |0\rangle$ of energy E_0 at time t_0, then the state at a future time differs from the initial state by a phase factor

$$H_0|\Psi_{t_0}\rangle = E_0|\Psi_{t_0}\rangle \implies |\Psi_t\rangle = e^{-i\frac{E_0}{\hbar}(t-t_0)}|\Psi_{t_0}\rangle. \tag{14.2}$$

This is a *stationary* state; if the quantum state started in an eigentstate, it remains in that eigenstate as long as there is no perturbation. But the eigen-state vector still 'rotates' in time with frequency $\omega_0 = E_0/\hbar$ in the Hilbert space as indicated schematically in Figure 14.1. It is called stationary because physical observables of the eigenstate will require not the amplitude, but the inner product, which is $\langle\Psi_t|\Psi_t\rangle = \langle\Psi_{t_0}|\Psi_{t_0}\rangle$. This is manifestly stationary in time.

Now let us perturb the system with a *time-dependent* term W_t. This perturbation can be due to a voltage applied on a semiconductor device, or electromagnetic waves (photons) incident on a semiconductor. The new Schrodinger equation for the time evolution of the state is

$$i\hbar\frac{\partial}{\partial t}|\Psi_t\rangle = [H_0 + W_t]|\Psi_t\rangle. \tag{14.3}$$

In principle, solving this equation will yield the complete future quantum states. In practice, this equation is unsolvable, even for the simplest of perturbations. Physically, the perturbation will 'scatter' a particle that was, say in state $|0\rangle$ to state $|n\rangle$. However, we had noted that even in the *absence* of perturbations, the eigen-state vectors were already

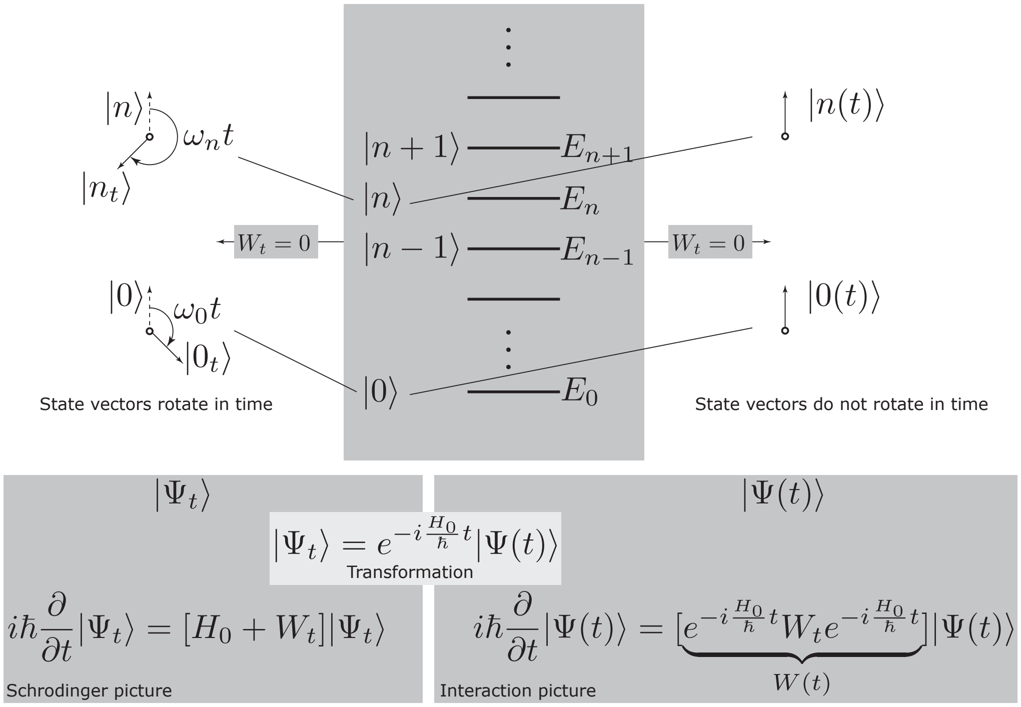

$|n\rangle$
$\omega_n t$
$|n_t\rangle$
$|0\rangle$
$\omega_0 t$
$|0_t\rangle$
State vectors rotate in time
$|n+1\rangle$
E_{n+1}
$|n\rangle$
E_n
$|n-1\rangle$
E_{n-1}
$|0\rangle$
E_0
$W_t = 0$
$W_t = 0$
$|n(t)\rangle$
$|0(t)\rangle$
State vectors do not rotate in time
$|\Psi_t\rangle$
$|\Psi(t)\rangle$
$|\Psi_t\rangle = e^{-i\frac{H_0}{\hbar}t}|\Psi(t)\rangle$
Transformation
$i\hbar\frac{\partial}{\partial t}|\Psi_t\rangle = [H_0 + W_t]|\Psi_t\rangle$
Schrodinger picture
$i\hbar\frac{\partial}{\partial t}|\Psi(t)\rangle = [\underbrace{e^{-i\frac{H_0}{\hbar}t}W_t e^{-i\frac{H_0}{\hbar}t}}_{W(t)}]|\Psi(t)\rangle$
Interaction picture

Fig. 14.1: Schrodinger vs. Interaction pictures of time-evolution of quantum state.

evolving with time in the Hilbert space. For example, state vector $|0\rangle$ was rotating at an angular frequency ω_0, and state vector $|n\rangle$ at ω_n. This is shown schematically in the left of Figure 14.1. It would be nice to work with *unperturbed* state vectors that do not change in time, as in the right of Figure 14.1. This calls for a transformation to a vector space that 'freezes' the time evolution of the unperturbed eigen state-vectors. Such a transformation is achieved by the relation

$$|\Psi_t\rangle = e^{-i\frac{H_0}{\hbar}t}|\Psi(t)\rangle, \tag{14.4}$$

where H_0 is the Hamiltonian *operator*. Note that the operator now sits in the exponential, but it should not worry us much. We will see that it is rather useful to have it up there. The reason for this non-obvious transformation is because when we put this into the Schrodinger equation in Equation 14.3, we get

$$i\hbar\left(-\frac{i}{\hbar}H_0 e^{-i\frac{H_0}{\hbar}t}|\Psi(t)\rangle + e^{-i\frac{H_0}{\hbar}t}\frac{\partial}{\partial t}|\Psi(t)\rangle\right) = [H_0 + W_t]e^{-i\frac{H_0}{\hbar}t}|\Psi(t)\rangle, \tag{14.5}$$

and there is a crucial cancellation, leaving us with

$$\boxed{i\hbar\frac{\partial}{\partial t}|\Psi(t)\rangle = [e^{+i\frac{H_0}{\hbar}t}W_t e^{-i\frac{H_0}{\hbar}t}]|\Psi(t)\rangle = W(t)|\Psi(t)\rangle} \tag{14.6}$$

where $W(t) = e^{+i\frac{H_0}{\hbar}t}W_t e^{-i\frac{H_0}{\hbar}t}$. Can we take the *operator* $e^{-i\frac{H_0}{\hbar}t}$ from the left to the right side as $e^{+i\frac{H_0}{\hbar}t}$? Yes we can, because $e^{+i\frac{H_0}{\hbar}t} \cdot e^{-i\frac{H_0}{\hbar}t} = I$, the identity operator.

The boxed form of the time-evolution is called the *interaction* picture, as opposed to the conventional form of Equation 14.3, which is called the 'Schrodinger' picture. Note that if there is no perturbation, $W_t = 0 \implies W(t) = 0 \implies i\hbar\frac{\partial|\Psi(t)\rangle}{\partial t} = 0$. Then, $|\Psi(t)\rangle = |\Psi(t_0)\rangle$, and we have managed to find the state vector representation in which the unperturbed eigenvectors are indeed frozen in time.

Now lets turn the perturbation W_t on. Formally, the state vector at time t in the interaction representation is obtained by integrating both sides:

$$|\Psi(t)\rangle = |\Psi(t_0)\rangle + \frac{1}{i\hbar}\int_{t_0}^{t} dt' W(t')|\Psi(t')\rangle, \tag{14.7}$$

and it looks as if we have solved the problem. However, there is a catch - the unknown state vector $|\Psi(t)\rangle$ appears also on the right side - inside the integral. This is also a recursive relation! It reminds of the Brilloiun-Wigner form of non-degenerate perturbation theory. Let's try to iterate the formal solution once:

$$|\Psi(t)\rangle = |\Psi(t_0)\rangle + \frac{1}{i\hbar}\int_{t_0}^{t} dt' W(t')\left[|\Psi(t_0)\rangle + \frac{1}{i\hbar}\int_{t_0}^{t'} dt'' W(t'')|\Psi(t'')\rangle\right], \tag{14.8}$$

and then keep going:

$$|\Psi(t)\rangle = \underbrace{|\Psi(t_0)\rangle}_{\sim W^0} + \underbrace{\frac{1}{i\hbar}\int_{t_0}^{t} dt' W(t')|\Psi(t_0)\rangle}_{\sim W^1} + \underbrace{\frac{1}{(i\hbar)^2}\int_{t_0}^{t} dt' W(t')\int_{t_0}^{t'} dt'' W(t'')|\Psi(t_0)\rangle}_{\sim W^2} + ... \tag{14.9}$$

We thus obtain a formal perturbation series to many orders. The hope is that the series converges rapidly if the perturbation is 'small', because successive terms increase as a power law, which for a small number gets even smaller. Let's accept that weak argument now at face value, and we return later to address, justify, and where possible, fix this cavalier approximation.

Let $|\Psi(t_0)\rangle = |0\rangle$ be the initial state of the quantum system. The perturbation is turned on at time t_0. The probability *amplitude* for the system to be found in state $|n\rangle$ at time $t(> t_0)$ is $\langle n|\Psi_t\rangle$. Note the Schrodinger representation! But the transformation from Schrodinger to interaction picture helps: $\langle n|\Psi_t\rangle = \langle n|e^{-i\frac{H_0}{\hbar}t}\Psi(t)\rangle = e^{-i\frac{E_n}{\hbar}t}\langle n|\Psi(t)\rangle$. This implies $|\langle n|\Psi_t\rangle|^2 = |\langle n|\Psi(t)\rangle|^2$ - for all eigenstates $|n\rangle$. Let us make an approximation in this section and retain only the first order term in the perturbation series. We will return later and discuss the higher order terms that capture multiple-scattering events. Retaining only the terms of Eq. 14.9 to first order in the perturbation W gives

$$\langle n|\Psi(t)\rangle \approx \underbrace{\langle n|0\rangle}_{=0} + \frac{1}{i\hbar}\int_{t_0}^{t} dt' \langle n|W(t')|0\rangle = \frac{1}{i\hbar}\int_{t_0}^{t} dt' \langle n|e^{+i\frac{H_0}{\hbar}t'} W_{t'} e^{-i\frac{H_0}{\hbar}t'}|0\rangle. \tag{14.10}$$

Let us assume the perturbation to be of the form $W_t = e^{\eta t}W$ representing a 'slow turn on', with $\eta = 0^+$, and $W = W(r)$ a function that depends only on space. If $\eta = 0$, then the perturbation is time-independent. But if $\eta = 0^+$, then $e^{\eta t_0} \to 0$ as $t_0 \to -\infty$. This construction thus effectively kills the perturbation far in the distant past, but slowly turns it on to full strength at $t = 0$. We will discuss more of the physics buried inside η later. For now, we accept it as a mathematical construction, with the understanding to take the limit $\eta \to 0$ at the end. Then, the amplitude in state $|n\rangle$ simplifies:

$$\langle n|\Psi(t)\rangle \approx \frac{1}{i\hbar}\int_{t_0}^{t} dt' \underbrace{\boxed{\langle n|e^{+i\frac{H_0}{\hbar}t'}}}_{e^{+i\frac{E_n}{\hbar}t'}\langle n|} e^{\eta t'} W \underbrace{\boxed{e^{-i\frac{H_0}{\hbar}t'}|0\rangle}}_{e^{-i\frac{E_0}{\hbar}t'}|0\rangle} = \frac{\langle n|W|0\rangle}{i\hbar}\int_{t_0}^{t} dt' e^{i\left(\frac{E_n - E_0}{\hbar}\right)t'} e^{\eta t'}, \tag{14.11}$$

and the integral over time may be evaluated exactly to yield

$$\int_{t_0}^{t} dt' e^{i(\frac{E_n-E_0}{\hbar})t'} e^{\eta t'} = \frac{e^{i(\frac{E_n-E_0}{\hbar})t}e^{\eta t} - e^{i(\frac{E_n-E_0}{\hbar})t_0}e^{\eta t_0}}{i\left(\frac{E_n-E_0}{\hbar}\right)+\eta} \underbrace{=}_{t_0\to-\infty} \frac{e^{i(\frac{E_n-E_0}{\hbar})t}e^{\eta t}}{i\left(\frac{E_n-E_0}{\hbar}\right)+\eta}. \tag{14.12}$$

The amplitude then is

$$\langle n|\Psi(t)\rangle \approx \frac{\langle n|W|0\rangle}{i\hbar} \cdot \frac{e^{i(\frac{E_n-E_0}{\hbar})t}e^{\eta t}}{i\left(\frac{E_n-E_0}{\hbar}\right)+\eta} = \langle n|W|0\rangle \cdot \frac{e^{i(\frac{E_n-E_0}{\hbar})t}e^{\eta t}}{(E_0-E_n)+i\hbar\eta}. \tag{14.13}$$

The *probability* of the state making a transition from $|0\rangle$ to $|n\rangle$ at time t is

$$|\langle n|\Psi_t\rangle|^2 = |\langle n|\Psi(t)\rangle|^2 \approx |\langle n|W|0\rangle|^2 \frac{e^{2\eta t}}{(E_0-E_n)^2+(\hbar\eta)^2}. \tag{14.14}$$

The *rate* of transitions from state $|0\rangle \to |n\rangle$ is

$$\frac{1}{\tau_{|0\rangle\to|n\rangle}} = \frac{d}{dt}|\langle n|\Psi(t)\rangle|^2 \approx |\langle n|W|0\rangle|^2 \left(\frac{2\eta}{(E_0-E_n)^2+(\hbar\eta)^2}\right) e^{2\eta t}. \tag{14.15}$$

Now we take $\eta \to 0^+$. The third term $e^{2\eta t} \to 1$, but we must be careful with the quantity in the bracket. When $\eta \to 0$, this quantity is 0, *except* when the term $E_0 - E_n = 0$; then the term seems indeterminate. By making a plot of this function, we can convince ourselves that it approaches a Dirac delta function in the variable $E_0 - E_n$. The mathematical identity $\lim_{\eta\to 0^+}\frac{2\eta}{x^2+\eta^2} = \lim_{\eta\to 0^+}\frac{1}{i}[\frac{1}{x-i\eta} - \frac{1}{x+i\eta}] = 2\pi\delta(x)$, where $\delta(...)$ confirms this: in the limit, the term indeed becomes the Dirac-delta function.

Then, using $\delta(ax) = \delta(x)/|a|$, the rate of transitions is given by

$$\boxed{\frac{1}{\tau_{|0\rangle\to|n\rangle}} \approx \frac{2\pi}{\hbar}|\langle n|W|0\rangle|^2\delta(E_0-E_n),} \tag{14.16}$$

which is the Fermi's golden rule. The general form is $2\pi/\hbar$ times the transition matrix element squared, times a Dirac-delta function as a statement of energy conservation.

14.3 Perturbations oscillating in time

Now suppose the perturbation potential was oscillating in time. We will encounter such perturbations frequently, in the form of electron-photon, or electron-phonon interactions. The mathematical nature of such perturbations with a slow turn-on is

$$W_t = 2We^{\eta t}\cos(\omega t) = e^{\eta t}W(e^{i\omega t} + e^{-i\omega t}) \tag{14.17}$$

which leads to a $|0\rangle \to |n\rangle$ transition amplitude

$$\langle n|\Psi(t)\rangle \approx \frac{\langle n|W|0\rangle}{i\hbar}\left(\int_{t_0}^{t} dt' e^{i(\frac{E_n-E_0+\hbar\omega}{\hbar})t'}e^{\eta t'} + \int_{t_0}^{t} dt' e^{i(\frac{E_n-E_0-\hbar\omega}{\hbar})t'}e^{\eta t'}\right), \tag{14.18}$$

Similar to Equations 14.12 and 14.13, evaluating the integral with $t_0 \to -\infty$, we get the amplitude for transitions

$$\langle n|\Psi(t)\rangle \approx \langle n|W|0\rangle \cdot \left(\frac{e^{i(\frac{E_n-E_0+\hbar\omega}{\hbar})t}e^{\eta t}}{(E_0-E_n+\hbar\omega)+i\hbar\eta} + \frac{e^{i(\frac{E_n-E_0-\hbar\omega}{\hbar})t}e^{\eta t}}{(E_0-E_n-\hbar\omega)+i\hbar\eta}\right). \tag{14.19}$$

The probability is then

$$\begin{aligned}|\langle n|\Psi(t)\rangle|^2 \approx |\langle n|W|0\rangle|^2 \cdot [\frac{e^{2\eta t}}{(E_0-E_n+\hbar\omega)^2+(\hbar\eta)^2} + \frac{e^{2\eta t}}{(E_0-E_n-\hbar\omega)^2+(\hbar\eta)^2} + \\ \frac{e^{2i\omega t}e^{2\eta t}}{(E_0-E_n+\hbar\omega+i\hbar\eta)(E_0-E_n-\hbar\omega-i\hbar\eta)} + \\ \frac{e^{-2i\omega t}e^{2\eta t}}{(E_0-E_n+\hbar\omega-i\hbar\eta)(E_0-E_n-\hbar\omega+i\hbar\eta)}]\end{aligned} \tag{14.20}$$

The rate of transition is then

$$\begin{aligned}\frac{d}{dt}|\langle n|\Psi(t)\rangle|^2 \approx |\langle n|W|0\rangle|^2 \cdot [\frac{2\eta e^{2\eta t}}{(E_0-E_n+\hbar\omega)^2+(\hbar\eta)^2} + \frac{2\eta e^{2\eta t}}{(E_0-E_n-\hbar\omega)^2+(\hbar\eta)^2} + \\ \frac{2(\eta+i\omega)e^{2i\omega t}e^{2\eta t}}{(E_0-E_n+\hbar\omega+i\hbar\eta)(E_0-E_n-\hbar\omega-i\hbar\eta)} + \\ \frac{2(\eta-i\omega)e^{-2i\omega t}e^{2\eta t}}{(E_0-E_n+\hbar\omega-i\hbar\eta)(E_0-E_n-\hbar\omega+i\hbar\eta)}].\end{aligned} \tag{14.21}$$

Notice that the last two (interference) terms are a complex conjugate pair, which they must be, because the rate of transition is real. The sum is then 2× the real part of either term. After some manipulations, one obtains

$$\begin{aligned}\frac{d}{dt}|\langle n|\Psi(t)\rangle|^2 \approx \\ \langle n|W|0\rangle|^2 e^{2\eta t} \cdot \left(\frac{2\eta}{(E_0-E_n+\hbar\omega)^2+(\hbar\eta)^2} + \frac{2\eta}{(E_0-E_n-\hbar\omega)^2+(\hbar\eta)^2}\right)[1-\cos(2\omega t)] + \\ 2\sin(2\omega t)\left(\frac{E_0-E_n+\hbar\omega}{(E_0-E_n+\hbar\omega)^2+(\hbar\eta)^2} - \frac{E_0-E_n-\hbar\omega}{(E_0-E_n-\hbar\omega)^2+(\hbar\eta)^2}\right).\end{aligned} \tag{14.22}$$

Note that the rate has a part that does not oscillate, and another which does, with *twice* the frequency of the perturbing potential. If we average over a few periods of the oscillation, $\langle\cos(2\omega t)\rangle_t = \langle\sin(2\omega t)\rangle_t = 0$. Then, by taking the limit $\eta \to 0^+$ in the same fashion as in Equation 14.16, we obtain the Fermi's golden rule for oscillating perturbations:

$$\boxed{\frac{1}{\tau_{|0\rangle\to|n\rangle}} \approx \frac{2\pi}{\hbar} \times |\langle n|W|0\rangle|^2 \times [\underbrace{\delta(E_0-E_n+\hbar\omega)}_{\text{absorption}} + \underbrace{\delta(E_0-E_n-\hbar\omega)}_{\text{emission}}].} \tag{14.23}$$

The Dirac-delta functions now indicate that the exchange of energy between the quantum system and the perturbing field is through *quanta* of energy: either by absorption, leading to $E_n = E_0 + \hbar\omega$, or emission, leading to $E_n = E_0 - \hbar\omega$. The rates of each individual processes are the same. Which process (emission or absorption) dominates depends on the occupation functions of the quantum states.

14.4 Transitions to a continuum of states

The Fermi golden rule results in Equation 14.16 and 14.23 are in a form suitable for tracking transitions between *discrete*, or individual states $|0\rangle$ and $|n\rangle$. For many situations encountered in semiconductors, these transitions will be between states within, or between energy bands, where a *continuum* of states exist. In those cases, the *net* transition rate will be obtained by summing over all relevant states. Even the transition between manifestly discrete states - for example the electron ground state of hydrogen atom to the first excited state - by the absorption of a photon - occurs by the interaction between the discrete electron states and the states of the electromagnetic spectrum, which forms a continuum.

As an example, consider the transitions between electron states in the conduction band due to a point scatterer in a 3D semiconductor. Let us say the point scatterer potential is $W(r) = V_0\delta(\mathbf{r})$, with V_0 in units of eV·m^3. This is not an oscillating potential, so we use the golden rule result of Equation 14.16. We first find the matrix element between electron states $|\mathbf{k}\rangle$ and $|\mathbf{k}'\rangle$:

$$\langle \mathbf{k}'|V_0\delta(\mathbf{r})|\mathbf{k}\rangle = \int d^3\mathbf{r}\left(\frac{e^{-i\mathbf{k}'\cdot\mathbf{r}}}{\sqrt{V}}\right)V_0\delta(\mathbf{r})\left(\frac{e^{+i\mathbf{k}\cdot\mathbf{r}}}{\sqrt{V}}\right) = \frac{V_0}{V}, \tag{14.24}$$

where we have used the property that the Fourier transform of a Dirac-delta function is equal to 1. Then, the transition (or scattering) rate to any state $|\mathbf{k}'\rangle$ is

$$\frac{1}{\tau(|\mathbf{k}\rangle \to |\mathbf{k}'\rangle)} = \frac{2\pi}{\hbar}\left(\frac{V_0}{V}\right)^2 \delta(E_{\mathbf{k}} - E_{\mathbf{k}'}). \tag{14.25}$$

The net scattering 'out' of state $|\mathbf{k}\rangle$ into the continuum of states $|\mathbf{k}'\rangle$ is then given by

$$\frac{1}{\tau(|\mathbf{k}\rangle)} = \sum_{\mathbf{k}'}\frac{1}{\tau(|\mathbf{k}\rangle \to |\mathbf{k}'\rangle)} = \frac{2\pi}{\hbar}\left(\frac{V_0}{V}\right)^2 \underbrace{\sum_{\mathbf{k}'}\delta(E_{\mathbf{k}} - E_{\mathbf{k}'})}_{D(E_{\mathbf{k}})}, \tag{14.26}$$

where we note that the sum over final states of the Dirac-delta function is the density of states $D(E_{\mathbf{k}})$ in units eV^{-1} of the electron at energy $E_{\mathbf{k}}$. This procedure illustrates an important result - the scattering rate for continuum of states is in general proportional to a density of states relevant to the problem. The strength of scattering increases as the square of the scattering potential. The occurrence of the (volume)^2 term in the denominator may be disconcerting at first. However, the macroscopic volume (or area, or length) terms will for most cases cancel out because of purely physical reasons. For example, for the problem illustrated here, if instead of just *one* point scatterer, we had N, the *density* of scatterers is $n_{sc} = N/V$. Together with the conversion process $\sum_{\mathbf{k}'} \to V\int d^3\mathbf{k}'/(2\pi)^3$, we obtain

$$\frac{1}{\tau(E_{\mathbf{k}})} = \frac{2\pi}{\hbar}\left(\frac{V_0}{V}\right)^2 n_{sc}V\int\frac{d^3\mathbf{k}'}{\frac{(2\pi)^3}{V}}\delta(E_{\mathbf{k}} - E_{\mathbf{k}'}) = \frac{2\pi}{\hbar}V_0^2 n_{sc}g(E_{\mathbf{k}}). \tag{14.27}$$

Here the density of states $g(E_{\mathbf{k}})$ is per unit volume, in units 1/(eV.m^3), as is standard in semiconductor physics. The scattering rate is linearly proportional to the density of scatterers. What is not immediately clear is how can we capture the effect of N scatterers by just multiplying the individual scatterer rate by N. This can be done if the scatterers are uncorrelated, as will be discussed in the transport chapters. For now, note that the macroscopic volume has canceled out, as promised.

14.5 Higher order transitions: Dyson series and Diagrams

In going from Equation 14.9 to 14.10, we had unceremoniously abandoned the higher order interaction terms. For most cases, this will serve us well. For the rest of the cases, here is how things can be made to work. Rewriting Equation 14.9 with the slow-turn on perturbation potential $W_t = e^{\eta t}W$, we get:

$$\langle n|\Psi(t)\rangle = \frac{1}{i\hbar}\int_{t_0}^{t} dt'\langle n|W(t')|0\rangle + \frac{1}{(i\hbar)^2}\int_{t_0}^{t} dt'\langle n|W(t')\int_{t_0}^{t'} dt''W(t'')|0\rangle + ... \tag{14.28}$$

In class, we used the *interaction representation* to write the perturbed quantum state at time t as $|\psi_t\rangle = e^{-i\frac{H_0}{\hbar}t}|\psi(t)\rangle$, where H_0 is the unperturbed Hamiltonian *operator*. This step helped us recast the time-dependent Schrodinger equation $i\hbar\frac{\partial}{\partial t}|\psi_t\rangle = (H_0 + W_t)|\psi_t\rangle$ to the simpler form $i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = W(t)|\psi(t)\rangle$, where $W(t) = e^{+i\frac{H_0}{\hbar}t}W_t e^{-i\frac{H_0}{\hbar}t}$ is the time-evolution operator. This equation was integrated over time to yield the Dyson series

$$\begin{aligned}|\psi(t)\rangle = \underbrace{|0\rangle}_{|\psi(t)\rangle^{(0)}} &+ \underbrace{\frac{1}{i\hbar}\int_{t_0}^{t} dt' W(t')|0\rangle}_{|\psi(t)\rangle^{(1)}} + \underbrace{\frac{1}{(i\hbar)^2}\int_{t_0}^{t} dt'\int_{t_0}^{t'} dt'' W(t')W(t'')|0\rangle}_{|\psi(t)\rangle^{(2)}} \\ &+ \underbrace{\frac{1}{(i\hbar)^3}\int_{t_0}^{t} dt'\int_{t_0}^{t'} dt''\int_{t_0}^{t''} dt''' W(t')W(t'')W(t''')|0\rangle}_{|\psi(t)\rangle^{(3)}} + ...,\end{aligned} \tag{14.29}$$

where $|\psi(t_0)\rangle = |0\rangle$ is the initial state. Restricting the Dyson series to the 1st order term in W for a perturbation of the the form $W_t = e^{\eta t}W(r)$, we derived Fermi's golden rule for the transition rate $\Gamma^{(1)}_{0\rightarrow n} = \frac{2\pi}{\hbar}|\langle n|W(r)|0\rangle|^2\delta(\epsilon_0 - \epsilon_n)$. We used the relation $\lim_{\eta\rightarrow 0^+}\frac{2\eta}{x^2+\eta^2} = 2\pi\delta(x)$ in this process.

The second and third order terms in W in the Dyson series lead to a modified golden rule result

$$\Gamma_{0\rightarrow n} = \frac{2\pi}{\hbar}|\langle n|W|0\rangle + \sum_m \frac{\langle n|W|m\rangle\langle m|W|0\rangle}{\epsilon_0 - \epsilon_m + i\eta\hbar} + \sum_{k,l}\frac{\langle n|W|k\rangle\langle k|W|l\rangle\langle l|V|0\rangle}{(\epsilon_0 - \epsilon_k + 2i\eta\hbar)(\epsilon_0 - \epsilon_l + i\eta\hbar)} + ...|^2\delta(\epsilon_0 - \epsilon_n), \tag{14.30}$$

where in the end we take $\eta \rightarrow 0^+$. We identify the Green's function propagators of the form $G = \sum_m \frac{|m\rangle\langle m|}{\epsilon_0 - \epsilon_m + i\eta\hbar}$. Thus, the result to higher orders may be written in the compact form

$$\Gamma_{0\rightarrow n} = \frac{2\pi}{\hbar}|\langle n|W + WGW + WGWGW + ...|0\rangle|^2\delta(\epsilon_0 - \epsilon_n). \tag{14.31}$$

'Feynman' diagrams[1] corresponding to the terms in the series can now be sketched for the problem, showing the *virtual* states explicitly for the higher order terms. (To be drawn.)

14.6 Fate of the initial state: Self Energy

Now for something as trivial as profound: time-independent perturbation theory is a limiting case of time-dependent perturbation! Well, well - we now see that we can recover all of time-independent perturbation theory such as the Rayleigh-Schrodinger and Brillouin-Wigner theories by extending the time-evolution operator.

What is the probability that at time t after turning on a perturbation V_t, we can still find the state in the *initial* state $|0\rangle$? We must find the amplitude $\langle 0|\psi_t\rangle$ and square

[1] More accurately, Goldstone diagrams.

it to get the answer. Doing so will lead to a deep connection between time-dependent and time-independent perturbation theories. In fact, it will show that time-independent perturbation theory is a *special case* of time-dependent perturbation theory - just as 'dc' is a special case of 'ac' with frequency $\omega \to 0$.

Since $\langle 0|\psi_t\rangle = e^{-i\epsilon_0 t/\hbar}\langle 0|\psi(t)\rangle$, we can write the time-dependent Schrodinger equation as

$$i\hbar\frac{\partial}{\partial t}\ln\langle 0|\psi(t)\rangle = \langle 0|V(t)|0\rangle + \sum_n{}^{'}\langle 0|V(t)|n\rangle\frac{\langle n|\psi(t)\rangle}{\langle 0|\psi(t)\rangle}, \tag{14.32}$$

where the prime over the sum indicates that $|n\rangle = |0\rangle$ is excluded. Let the perturbation be $V_t = e^{\eta t}V$.

Now using the Dyson series to the 1st order in V, in the limit $\eta \to 0^+$, we get

$$i\hbar\frac{\partial}{\partial t}\ln\langle 0|\psi(t)\rangle = \langle 0|V|0\rangle + \sum_n{}^{'}\frac{|\langle n|V|0\rangle|^2}{\epsilon_0 - \epsilon_n + i\eta\hbar} \tag{14.33}$$

Notice that the terms on the right side look very similar to the Rayleigh-Schrodinger perturbation in energy to first and second order in V.

Now by integration, the required probability amplitude is

$$\boxed{\langle 0|\psi_t\rangle \sim e^{-i\frac{E_0}{\hbar}t}e^{-\frac{\Gamma}{2}t}} \text{ where } \boxed{E_0 = \epsilon_0 + \langle 0|V|0\rangle + P\sum_n\frac{|\langle n|V|0\rangle|^2}{\epsilon_0 - \epsilon_n}} \text{ and } \boxed{\Gamma = \sum_n\frac{2\pi}{\hbar}|\langle n|V|0\rangle|^2\delta(\epsilon_0 - \epsilon_n)} \tag{14.34}$$

Here $P(...)$ is the principal part. We see that the time-independent perturbation theory may be recovered as a special case of time-dependent perturbation theory. The 'imaginary' component of energy $\sum = i\Gamma/2$ is called the *self-energy* of state $|0\rangle$. Note that the probability is conserved at all times : i.e., $\sum|\langle n|\psi_t\rangle|^2 = 1$.

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 15

The Boltzmann Transport Equation

Contents

15.1 Introduction 147
15.2 The Boltzmann Transport Equation 147
15.3 Scattering in equilibrium 149
15.4 Scattering in steady state 150
15.5 Formal Boltzmann Transport theory 151

15.1 Introduction

When a crystal is in *equilibrium*[1], electrons in it are distributed in energy according to the Fermi-Dirac function

$$f_0(E) = \frac{1}{1 + e^{\frac{E-E_F}{kT}}}, \tag{15.1}$$

where E is the *total* energy of the electron, and E_F represents the Fermi energy. The temperature T is the same for the atomic lattice and the electrons, because they are in equilibrium. The total energy, for example, for electrons in the conduction band of a semiconductor is the sum of the band-edge potential energy $E_c(r)$ and the kinetic energy given by the conduction bandstructure $\mathcal{E}_c(k)$

$$E = E_c(r) + \mathcal{E}_c(k) = E_c(x) + \frac{\hbar^2 k^2}{2m_c^\star} \implies f_0(x,k) = \frac{1}{1 + e^{\frac{E_c(x)+\mathcal{E}_c(k)-E_F}{kT}}}, \tag{15.2}$$

where we consider a 1D case for the initial discussion ($r = x, k = k_x$). As a result, the equilibrium Fermi-Dirac occupation function depends on the real space and the k-space, and may be formally written as $f_0(x,k)$. This equilibrium function does not depend on time. The value of the function lies in $0 \leq f_0(x,k) \leq 1$, physically representing the occupation probability of finding an electron in real space location x, and in the energy eigenstate $\mathcal{E}_c(k)$. If the crystal was sitting at a higher or lower temperature, the T changes, but the system is still in equilibrium: the occupation function of states continues to be $f_0(x,k)$. What happens to the occupation function when the state of equilibrium is disturbed? That is what the Boltzmann transport equation tells us.

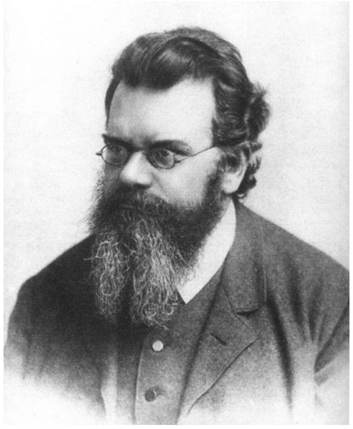


Fig. 15.1: Ludwig Boltzmann, the father of statistical mechanics and kinetic theory of gases. He discovered the formula for entropy or disorder as a mathematical concept: this formula $S = k_B \log\Omega$ is inscribed on his gravestone. The concept of entropy permeates all branches of sciences, including communication systems. Boltzmann ended his life by hanging himself in 1906.

15.2 The Boltzmann Transport Equation

Consider now the situation where a voltage has been applied across contacts made to a semiconductor. The electrons respond to the electric field by changing their (x,k) coordinates as indicated in the left of Figure 15.3. The (x,k) space is referred to as the 'phase-space', a term borrowed from classical mechanics. In classical mechanics, the motion,

[1] *Equilibrium* is not the same as *steady state*, as will be clarified.

or *trajectory* of a particle in the phase space is *completely* determined by Newton's laws[2]. But the fundamental axiom of quantum mechanics forbids the simultaneous determination of (x, k) by the uncertainty relation $\Delta x \cdot \Delta k \sim 1$. The location of the 'particle' is thus diffuse in the phase space, as indicated in Figure 15.3.

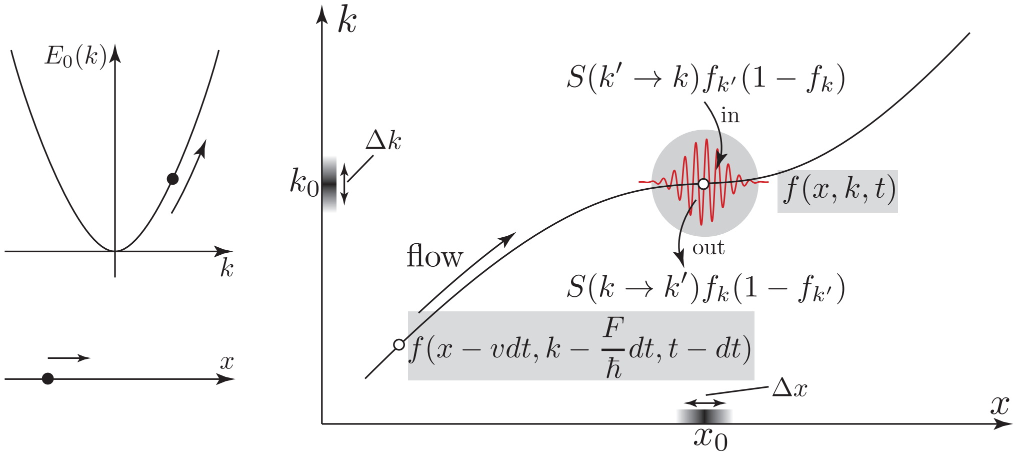

$E_0(k)$
k
x
k
Δk
k_0
$S(k' \to k) f_{k'}(1 - f_k)$
in
$f(x, k, t)$
out
$S(k \to k') f_k(1 - f_{k'})$
flow
$f(x - vdt, k - \frac{F}{\hbar}dt, t - dt)$
Δx
x
x_0

Fig. 15.2: Scattering term of Boltzmann transport equation depicting the inflow and outflow of the distribution function.

But we have introduced the concept of a *wavepacket* for electrons precisely to deal with this situation. If we agree to represent the electrons in the *effective mass approximation* developed in Chapter 11, we can track the trajectory of the *center* of the wavepacket representing the electron, as indicated in Figure 15.3. Instead of tracking the location of the wavepacket, it will be more fruitful to track the *occupation function* of its center, which we denote as $f(x, k, t)$. This occupation function has the same physical meaning as the Fermi-Dirac function, except it is also the probability of occupation under *non-equilibrium* conditions. In other words, if we remove the perturbation (for example the external voltage), and give the system enough time to relax, $f(x, k, t) \to f_0(x, k)$ as $t \to \infty$.

From Figure 15.3, the occupation function at time t at (x, k) must be identical to the value at $t - dt$ centered at $(x - vdt, k - \frac{F}{\hbar}dt)$, where v is the *group velocity* of the wavepacket, and $F = \hbar \frac{dk}{dt}$ governs the translation of k due to force F. This statement is true for *ballistic* motion, for which the wavepackets traverse well-defined trajectories. The occupation probability flows with the wavepacket. The increase at a point in phase space is at the expense of the decrease of another. This is another statement of the continuity equation. Then, we can write $f(x, k, t) = f(x - vdt, k - \frac{F}{\hbar}dt, t - dt)$. If there are random *scattering* events however, we must also allow for $f(x, k, t)$ to *increase* by scattering 'in' from nearby k' points on other trajectories, and for it to *decrease* by scattering 'out'. Then, the continuity equation reads

$$f = f(x, k, t) = f(x - vdt, k - \frac{F}{\hbar}dt, t - dt) + (S_{in} - S_{out})dt, \tag{15.3}$$

where S_{in} and S_{out} are the in- and out-scattering *rates*, in units of 1/s. Taylor-expanding the first term on the right and rearranging, we get the 1D Boltzmann transport equation

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} + \frac{F}{\hbar}\frac{\partial f}{\partial k} = S_{in} - S_{out}. \tag{15.4}$$

Generalizing to higher dimensions, we write

[2] Newton's laws give the least-action paths obtained by minimizing the action in Lagrangian/Hamiltonian classical mechanics

$$\frac{\partial f}{\partial t} + \mathbf{v}_k \cdot \nabla_r f + \frac{\mathbf{F}}{\hbar} \cdot \nabla_k f = S_{in} - S_{out}. \tag{15.5}$$

Writing $f(r,k,t) = f_k$, it is clear from Figure 15.3 that the net in- and out-scattering rates depend on the occupation functions of the initial and final states. Let $S(k \to k')$ be the scattering rate from state $|k\rangle \to |k'\rangle$. Enforcing Pauli exclusion principle for electrons, we must have

$$S_{in} = S(k' \to k) f_{k'}(1 - f_k), \tag{15.6}$$

$$S_{out} = S(k \to k') f_k (1 - f_{k'}). \tag{15.7}$$

Now to track the occupation function of state $|k\rangle$, we must allow for in- and out-scattering to *all* other states $|k'\rangle$. Then, the complete Boltzmann transport equation reads:

$$\boxed{\frac{\partial f_k}{\partial t} + \mathbf{v}_k \cdot \nabla_r f_k + \frac{\mathbf{F}}{\hbar} \cdot \nabla_k f_k = \underbrace{\sum_{k'} [S(k' \to k) f_{k'}(1 - f_k) - S(k \to k') f_k (1 - f_{k'})]}_{\text{scattering term, } \hat{C} f_k}.} \tag{15.8}$$

Each scattering event $S(k \to k') = \frac{2\pi}{\hbar} |W_{k,k'}|^2 \delta(E_k - E_{k'} \pm \hbar\omega)$ is described by the Fermi's golden rule and used to explain the optical properties of semiconductors. Solving the BTE in Equation 15.8 will yield $f(r,k,t)$, from which the charge (or heat, spin, ...) transport properties of the electron distribution can be obtained. The formidable appearance of the equation is because of its generality. Let's break it into steps, by first considering the so-called 'collision integral' $\hat{C} f_k$, or the scattering term on the right.

15.3 Scattering in equilibrium

Do random scattering events occur when the system is unperturbed, i.e., in equilibrium? You bet! Scattering is what enables the system to attain equilibrium in the first place. When the electron system is in equilibrium with say the lattice at temperature T, the rate of *every* scattering event is *exaclty* counterbalanced by the reverse process. This goes by the fancy name of the 'principle of detailed balance', introduced by Boltzmann himself. This requires the RHS of Equation 15.8 with $f_k \to f_{0k} = 1/(1 + e^{(E_k - E_F)/kT})$ at equilibrium to follow not just $\hat{C} f_k = 0$, but *every term* in the sum to be zero:

$$S(k' \to k) f_{0k'}(1 - f_{0k}) = S(k \to k') f_{0k}(1 - f_{0k'}), \tag{15.9}$$

which requires

$$\frac{S(k' \to k)}{S(k \to k')} = \frac{1 - f_{0k'}}{f_{0k'}} \cdot \frac{f_{0k}}{1 - f_{0k}} = e^{\frac{E_{k'} - E_k}{kT}}. \tag{15.10}$$

Enforcing the principle of detailed balance is telling us that for electrons, the scattering rate from state $|k\rangle \to |k'\rangle$ is *not the same* as for the reverse process, *unless* the energies of the two states are the same. For *elastic* scattering events $E_k = E_{k'}$ for which the energy of the electron is unchanged, the scattering rate $S(k \to k') = S(k' \to k)$ is the same for a process and its reverse. But for inelastic scattering events with $E_{k'} - E_k = \hbar\omega$, the scattering rate going uphill in energy is slower: $S(k \to k') = S(k' \to k) e^{-\hbar\omega/kT}$. The scattering rates $S(...)$ remain the same whether electrons are in equilibrium or not, the occupation functions f are what change.

Consider for example, the electron scattering rate due to either the absorption or emission of phonons of energy $\hbar\omega$. The rate of phonon absorption must be proportional to the number of phonons already present, i.e, $S_{abs} \propto n_{ph}$. The rate of phonon emission by an electron requires it to go downhill in energy, thus $S_{em} = S_{abs}e^{\hbar\omega/kT} \propto e^{\hbar\omega/kT}n_{ph}$. Since the number of phonons in mode ω is given by the Bose-Einstein function $n_{ph} = 1/(e^{\hbar\omega/kT} - 1)$, we note that $e^{\hbar\omega/kT}n_{ph} = 1 + n_{ph}$. Thus, $S_{abs} \propto n_{ph}$, but $S_{em} \propto (1 + n_{ph})$. Electrons are free to 'emit' phonons even when there are no phonons present - thus, the '1' represents *spontaneous* emission. But if there already are phonons present, the emission rate is enhanced, or *stimulated*; this is the reason for the $1 + n_{ph}$ proportionality of the net emission rate.

Note that nowhere have we explicitly needed that the quantum of energy $\hbar\omega$ be from a phonon - it could be from a photon, or other bosonic quanta. In other words, the concepts of absorption, spontaneous emission, and stimulated emission appear whenever fermionic (e.g. electron) systems are let to interact with bosonic (e.g. photons, phonons) systems. Such interactions play a critical role in electron transport in crystals as well as in electron-photon or light-matter interactions.

Can there be Fermi level or temperature gradients in equilibrium? To answer this question, we go back to the full BTE and noticing $\frac{\partial f_0}{\partial t} = 0$ at equilibrium, get

$$\mathbf{v}_k \cdot \nabla_r f_0 + \frac{\mathbf{F}}{\hbar} \cdot \nabla_k f_0 = 0. \tag{15.11}$$

Because $f_0 = \frac{1}{1+e^g}$ where $g(r,k,T) = \frac{E_c(r)+\mathcal{E}_c(k)-E_F(r)}{kT}$, we note that $\frac{\partial f_0}{\partial \mathcal{E}} = \frac{\partial g}{\partial \mathcal{E}}\frac{\partial f_0}{\partial g} = -\frac{1}{kT}\frac{e^g}{(1+e^g)^2} \implies \frac{\partial f_0}{\partial g} = kT\frac{\partial f_0}{\partial \mathcal{E}}$, and use the identities $\nabla_r f_0 = kT\frac{\partial f_0}{\partial \mathcal{E}}\nabla_r g$ and $\nabla_k f_0 = kT\frac{\partial f_0}{\partial \mathcal{E}}\nabla_k g$ with $\nabla_k g = \frac{\nabla_k \mathcal{E}}{kT} = \frac{\hbar \mathbf{v}_k}{kT}$, to rewrite the BTE as

$$kT \cdot \frac{\partial f_0}{\partial \mathcal{E}} \cdot \mathbf{v}_k \cdot [\frac{\mathbf{F}}{kT} + \nabla_r g] = 0. \tag{15.12}$$

We are writing $T_L(r) = T$ as a lattice temperature that potentially varies in space, and electrons are in equilibrium with this lattice temperature. We have allowed for the possibility for the Fermi level to vary with position. We will let the BTE tell us if these quantities *actually* do vary in space. Equilibrium requires the absence of *external* electric fields, but it does not rule out the presence of *internal* electric fields! Let $\mathbf{F} = \mathbf{F}(r)$ be an internal spatially varying field. From equation 15.12, $\frac{\mathbf{F}}{kT} + \nabla_r g = 0$ requires

$$\frac{1}{kT}\Big(\mathbf{F} + \nabla_r E_c(r) - \nabla_r E_F(r)\Big) + [E_c(r) + \mathcal{E}_c(k) - E_F(r)]\nabla_r(\frac{1}{kT}) = 0. \tag{15.13}$$

Now since $\mathbf{F} = -\nabla_r E_c(r)$, the first two terms in the large left bracket cancel and leave

$$-\nabla_r E_F(r) + [E_c(r) + \mathcal{E}_c(k) - E_F(r)]T\nabla_r(\frac{1}{T}) = 0. \tag{15.14}$$

15.4 Scattering in steady state

Steady state is different from equilibrium. $\hat{C}f = 0$ at equilbrium, but at steady state $\hat{C}f \neq 0$, and is precisely the restoring term on the right that holds the distribution function from drifting away from its equilibrium Fermi-Dirac value to infinity due to the driving forces $f_0 \to f$. We will see shortly that for weak driving forces such as small fields, or weak concentration or temperature gradients, the distribution function does not wander too far from equilibrium. Under certain approximations, the deviation is only linear, meaning

$$\hat{C}f(k) \approx -\frac{f(k) - f_0(k)}{\tau(k)}, \tag{15.15}$$

where $\tau(k)$ is a characteristic scattering (or relaxation) time, giving this approximate form of the collision integral the name 'Relaxation Time Approximation' or RTA.

Using Equation 15.10, the scattering term in the BTE of Equation 15.8 may be then written as

$$\hat{C}f_k = \sum_{k'} S(k \to k')\big(e^{\frac{E_{k'}-E_k}{kT}} f_{k'}(1-f_k) - f_k(1-f_{k'})\big). \tag{15.16}$$

We now consider two cases that will prove useful in understanding the electron transport properties in semiconductors. The first is *inelastic* scattering by optical phonons whose energy $E_{k'} - E_k = \hbar\omega_0 >> kT$. For such scattering events, the scattering term simplifies to

$$\hat{C}f_k \approx e^{\frac{\hbar\omega_0}{kT}} \sum_{k'} S(k \to k')f_{k'}, \tag{15.17}$$

where we assume $f_k f_{k'} << f_{k'}$. We will need this relation when we treat polar-optical phonon scattering in wide-bandgap semiconductors for which $\hbar\omega_0 >> kT$ holds.

The second situation is for *elastic* scattering, when $E_{k'} = E_k$. For such processes, the scattering term is exactly

$$\hat{C}f_k = \sum_{k'} S(k \to k')(f_{k'} - f_k) = \sum_{k'} S(k \to k')f_{k'} - \frac{f_k}{\tau(k)}. \tag{15.18}$$

Note that the second term on the right is linear in the unknown f_k for which we are solving the BTE. It appears with a characteristic scattering rate $\frac{1}{\tau(k)} = \sum_{k'} S(k \to k')$, which sums the scattering rate from $|k\rangle$ to all possible $|k'\rangle$.

15.5 Formal Boltzmann Transport theory

Much of the following summary is collected from textbooks and research articles. The main references for this section are Seeger [2], Wolfe et. al. [3], and Davies [4]. No claim to originality is made for much of the material. The subsection on *generalization* of mobility expressions for arbitrary dimensions and arbitrary degeneracy is original, though much of it is inspired from the references.

A distribution-function $f(\mathbf{k}, \mathbf{r}, t)$ is the probability of occupation of an electron at time t at $\mathbf{r}$ with wavevectors lying between $\mathbf{k}, \mathbf{k} + \mathbf{dk}$. Under equilibrium ($\mathbf{E} = \mathbf{B} = \nabla_r f = \nabla_T f = 0$, i.e., no external electric ($\mathbf{E}$) or magnetic ($\mathbf{B}$) field and no spatial and thermal gradients), the distribution function is found from quantum-statistical analysis to be given by the Fermi-Dirac function for fermions -

$$f_0(\varepsilon) = \frac{1}{1 + e^{\frac{\varepsilon_{\mathbf{k}} - \mu}{k_B T}}}, \tag{15.19}$$

where $\varepsilon_{\mathbf{k}}$ is the energy of the electron, μ is the Fermi energy, and k_B is the Boltzmann constant.

Any external perturbation drives the distribution function away from the equilibrium; the Boltzmann-transport equation (BTE) governs the shift of the distribution function from equilibrium. It may be written formally as [3]

$$\frac{df}{dt} = \frac{\mathbf{F_t}}{\hbar} \cdot \nabla_{\mathbf{k}} f(\mathbf{k}) + \mathbf{v} \cdot \nabla_{\mathbf{r}} f(\mathbf{k}) + \frac{\partial \mathbf{f}}{\partial \mathbf{t}}, \tag{15.20}$$

where on the right hand side, the first term reflects the change in distribution function due to the total field force $\mathbf{F_t} = \mathbf{E} + \mathbf{v} \times \mathbf{B}$, the second term is the change due to concentration gradients, and the last term is the local change in the distribution function. Since the total number of carriers in the crystal is constant, the total rate of change of

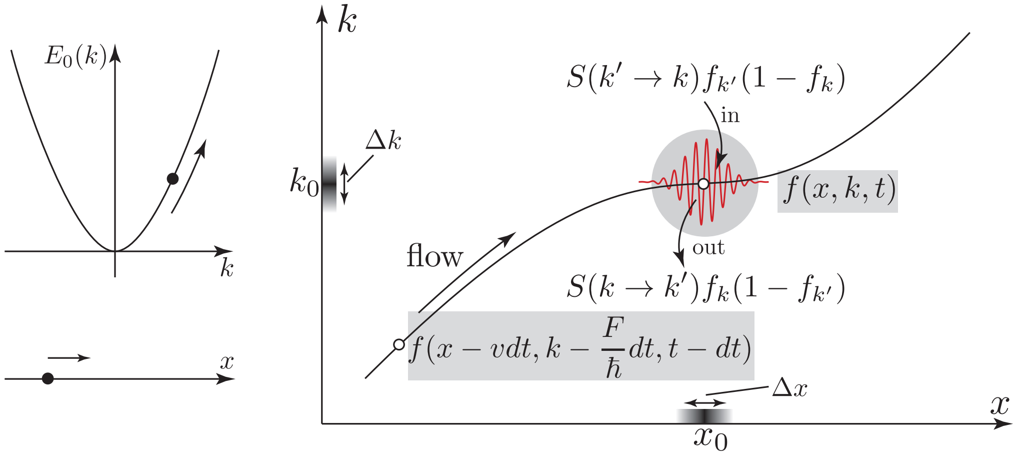

$E_0(k)$
k
x
k
Δk
k_0
$S(k' \to k) f_{k'}(1-f_k)$
in
$f(x,k,t)$
out
$S(k \to k') f_k(1-f_{k'})$
flow
$f(x - vdt, k - \frac{F}{\hbar}dt, t - dt)$
Δx
x
x_0

Fig. 15.3: Scattering term of Boltzmann transport equation depicting the inflow and outflow of the distribution function.

the distribution is identically zero by Liouville's theorem. Hence the *local* change in the distribution function is written as

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial t}|_{coll} - \frac{\mathbf{F_t}}{\hbar} \cdot \nabla_{\mathbf{k}} f(\mathbf{k}) - \mathbf{v} \cdot \nabla_{\mathbf{r}} \mathbf{f}(\mathbf{k}) + \frac{\partial \mathbf{f}}{\partial \mathbf{t}}, \tag{15.21}$$

where the first term has been split off from the field term since collision effects are not easily described by fields. The second term is due to applied field only and the third is due to concentration gradients.

Denoting the scattering rate from state $\mathbf{k} \to \mathbf{k}'$ as $S(\mathbf{k}, \mathbf{k}')$, the collision term is given by

$$\frac{\partial f(\mathbf{k})}{\partial t}|_{coll} = \sum_{\mathbf{k}'} [S(\mathbf{k}', \mathbf{k}) f(\mathbf{k}')[1 - f(\mathbf{k})] - S(\mathbf{k}, \mathbf{k}') f(\mathbf{k})[1 - f(\mathbf{k}')]]. \tag{15.22}$$

Figure 15.3 provides a visual representation of the scattering processes that form the collision term. The increase of the distribution function in the small volume $\Delta\mathbf{k}$ by particles flowing in by the field term is balanced by the net flow out by the two collision terms.

At equilibrium ($f = f_0$), the 'principle of detailed balance' enforces the condition

$$S(\mathbf{k}', \mathbf{k}) f_0(\mathbf{k}')[1 - f_0(\mathbf{k})] = S(\mathbf{k}, \mathbf{k}') f_0(\mathbf{k})[1 - f_0(\mathbf{k}')], \tag{15.23}$$

which translates to

$$S(\mathbf{k}', \mathbf{k}) e^{\frac{\varepsilon_{\mathbf{k}}}{k_B T}} = S(\mathbf{k}, \mathbf{k}') e^{\frac{\varepsilon_{\mathbf{k}'}}{k_B T}}. \tag{15.24}$$

In the special case of *elastic* scattering, $\varepsilon_{\mathbf{k}} = \varepsilon_{\mathbf{k}'}$, and as a result, $S(\mathbf{k}', \mathbf{k}) = S(\mathbf{k}, \mathbf{k}')$ irrespective of the nature of the distribution function. Using this, one rewrites the collision term as

$$\frac{\partial f(\mathbf{k})}{\partial t}|_{coll} = \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}')(f(\mathbf{k}') - \mathbf{f}(\mathbf{k})). \tag{15.25}$$

One can rewrite this collision equation as

$$\frac{df(\mathbf{k})}{dt} + \frac{f(\mathbf{k})}{\tau_q(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}') f(\mathbf{k}'), \tag{15.26}$$

where the *quantum scattering time* is defined as

$$\frac{1}{\tau_q(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k},\mathbf{k}'). \tag{15.27}$$

A particle prepared in state $|\mathbf{k}\rangle$ at time $t = 0$ by an external perturbation will be scattered into other states $|\mathbf{k}'\rangle$ due to collisions, and the distribution function in that state will approach the equilibrium distribution exponentially fast with the time constant $\tau_q(\mathbf{k})$ upon the removal of the applied field. The quantum scattering time $\tau_q(\mathbf{k})$ may be viewed as a 'lifetime' of the particle in the state $|\mathbf{k}\rangle$.

Let us now assume that the external fields and gradients have been turned on for a long time. They have driven the distribution function to a *steady state* value f from f_0. The perturbation is assumed to be small, i.e., distribution function is assumed not to deviate far from its equilibrium value of f_0. Under this condition, it is common practice to assume that

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial t}|_{coll} = -\frac{f - f_0}{\tau}, \tag{15.28}$$

where τ is a time scale characterizing the relaxation of the distribution. This is the relaxation time approximation, which is crucial for getting a solution of the Boltzmann transport equation.

When the distribution function reaches a steady state, the Boltzmann transport equation may be written as

$$\frac{\partial f}{\partial t} = -\left(\frac{f - f_0}{\tau}\right) - \frac{\mathbf{F_t}}{\hbar} \cdot \nabla_{\mathbf{k}} f(\mathbf{k}) - \mathbf{v} \cdot \nabla_{\mathbf{r}} \mathbf{f}(\mathbf{k}) = \mathbf{0}, \tag{15.29}$$

where the relaxation time approximation to the collision term has been used. In the absence of any concentration gradients, the distribution function is given by

$$f(\mathbf{k}) = f_0(\mathbf{k}) - \tau\frac{\mathbf{F_t}}{\hbar} \cdot \nabla_{\mathbf{k}} f. \tag{15.30}$$

Using the definition of the velocity $\mathbf{v} = 1/\hbar(\partial\varepsilon_{\mathbf{k}}/\partial k)$, the distribution function becomes

$$f(\mathbf{k}) = f_0(\mathbf{k}) - \tau\mathbf{F_t} \cdot \mathbf{v}\frac{\partial f(\mathbf{k})}{\partial\varepsilon}, \tag{15.31}$$

and since the distribution function is assumed to be close to f_0, we can make the replacement $f(\mathbf{k}) \to f_0(\mathbf{k})$, whence the distribution function

$$f(\mathbf{k}) = f_0(\mathbf{k}) - \tau\mathbf{F_t} \cdot \mathbf{v}\frac{\partial f_0(\mathbf{k})}{\partial\varepsilon} \tag{15.32}$$

is the *solution* of BTE for a perturbing force $\mathbf{F_t}$.

The external force $\mathbf{F_t}$ may be due to electric or magnetic fields. We first look for the solution in the presence of only the electric field; thus, $\mathbf{F_t} = -e\mathbf{E}$.

Using Equation 15.32, for elastic scattering processes one immediately obtains

$$f(\mathbf{k}') - f(\mathbf{k}) = \underbrace{e\tau\frac{\partial f_0}{\partial\varepsilon}\mathbf{E} \cdot \mathbf{v}}_{f(\mathbf{k})-f_0(\mathbf{k})}(1 - \frac{\mathbf{E} \cdot \mathbf{v}'}{\mathbf{E} \cdot \mathbf{v}}) \tag{15.33}$$

for a parabolic bandstructure ($\mathbf{v} = \hbar\mathbf{k}/m^\star$). Using this relation, the collision term in the form of the relaxation time approximation becomes

$$\frac{\partial f(\mathbf{k})}{\partial t} = \sum_{\mathbf{k}'} S(\mathbf{k},\mathbf{k}')(f(\mathbf{k}') - f(\mathbf{k})) = -\frac{(f(\mathbf{k}) - f_0(\mathbf{k}))}{\tau_m(\mathbf{k})}, \tag{15.34}$$

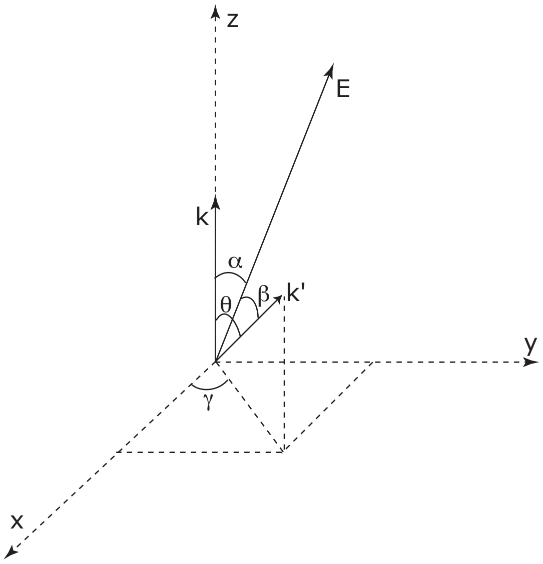


Fig. 15.4: Angular relations between the vectors in the Boltzmann transport equation.

where a new relaxation time is defined by

$$\frac{1}{\tau_m(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k},\mathbf{k}')(1 - \frac{\mathbf{E}\cdot\mathbf{k}'}{\mathbf{E}\cdot\mathbf{k}}). \tag{15.35}$$

This is the *momentum relaxation time.*

Let the vectors $\mathbf{k}, \mathbf{k}', \mathbf{E}$ be directed along random directions in the $3-$dimensional space. We fix the $z-$axis along $\mathbf{k}$ and the $y-$axis so that $\mathbf{E}$ lies in the $y-z$ plane. From Figure 15.4, we get the relation

$$\frac{\mathbf{k}'\cdot\mathbf{E}}{\mathbf{k}\cdot\mathbf{E}} = \cos\theta + \sin\theta\sin\gamma\tan\alpha, \tag{15.36}$$

where the angles are shown in the figure.

When the sum over *all* $\mathbf{k}'$ is performed for the collision term, the $\sin(\gamma)$ sums to zero and the momentum relaxation time $\tau_m(\mathbf{k})$ becomes

$$\frac{1}{\tau_m(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k},\mathbf{k}')(1 - \cos\theta). \tag{15.37}$$

We note here that this relation can be generalized to an arbitrary number of dimensions, the three-dimensional case was used as a tool. This is the general form for momentum scattering time, which is used heavily in the text for finding scattering rates determining mobility. It is related to mobility by the Drude relation $\mu = e\langle\tau(\mathbf{k})\rangle/m^{\star}$, where the momentum scattering time has been averaged over all energies of carriers.

The quantum scattering rate $1/\tau_q(\mathbf{k}) = \sum_{\mathbf{k}'} S(\mathbf{k},\mathbf{k}')$ and the momentum scattering rate $1/\tau_m(\mathbf{k}) = \sum_{\mathbf{k}'} S(\mathbf{k},\mathbf{k}')(1-\cos\theta)$ are both experimentally accessible quantities, and provide a valuable method to identify the nature of scattering mechanisms. The momentum scattering time $\tau_m(\mathbf{k})$ measures the average time spent by the particle moving along the external field. It differs from the quantum lifetime due to the $\cos\theta$ term. The angle θ is identified from Figure 15.4 as the angle between the initial and final wavevectors upon a scattering event. Thus for scattering processes that are isotropic $S(\mathbf{k},\mathbf{k}')$ has no angle dependence, the $\cos\theta$ term sums to zero, and $\tau_q = \tau_m$. However, for scattering processes that favor small angle ($\theta \to 0$) scattering, it is easily seen that $\tau_m > \tau_q$.

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 16

Maxwell Equations in a Nutshell

Contents

16.1 Introduction 155
16.2 Maxwell's equations . . . 156
16.3 Light emerges from Maxwell's equations 157
16.4 Maxwell's equations in $(\mathbf{k}, \omega)$ space 158
16.5 Maxwell's equations in material media 160
16.6 Need for a quantum theory of light 162

16.1 Introduction

Light has fascinated us for ages. And deservedly so. Everything we know about the earth and the universe is because of light. Light from the sun sustains life on earth. Learning to measure and understand the contents of light has enabled us to understand the origins of the universe in the big bang, and talk about its future. And one cannot forget the sheer visual pleasure of a beautiful sunset, a coral reef, or an iridescent flower in full blossom. Indeed, the beauty of light and color is a rare thing that scientists and artists agree to share and appreciate.

Fig. 16.1: James Clerk Maxwell

Our fascination with light has led to three of the greatest revolutions in 19th and 20th century physics. Sunlight used to be considered a 'gift of the Gods' and the purest indivisible substance, till Newton observed that passing it through a prism split it into multiple colors. Passing each of the colors through another prism could not split it further. Newton surmised that light was composed of particles, but in the early 19th century, Young proved that light was a wave because it exhibited interference and diffraction. Michael Faraday had a strong hunch that light was composed of a mixture of electric and magnetic fields, but could not back it up mathematically. The race for understanding the fabric of light reached a milestone when Maxwell gave Faraday's hunch a rigorous mathematical grounding. Maxwell's theory combined in one stroke electricity, magnetism, and light into an eternal braid[1]. The Maxwell equations predict the existence of light as a propagating electromagnetic wave. With Maxwell's **electromagnetic theory**, the 'cat' was out of the hat for light.

The second and third revolutions born out of light occurred in early 20th century in parallel. Trying to understand blackbody radiation, photoelectric effect, and the spectral lines of hydrogen atoms lead to the uncovering of **quantum mechanics**. And Einstein's fascination with the interplay of light and matter, of space and time led to the theory of **relativity**. Much of modern physics rests on these three pillars of light: that of electromagnetism, quantum mechanics, and relativity. It would be foolhardy to think that we know all there is to know about light. It will continue to amaze us and help probe deeper into the fabric of nature through similar revolutions in the future. In this chapter, we discuss Maxwell's theory of electromagnetism in preparation for the quantum picture, which is covered in the next chapter.

[1] J. R. Pierce famously wrote "To anyone who is motivated by anything beyond the most narrowly practical, it is worthwhile to understand Maxwell's equations simply for the good of his soul."

16.2 Maxwell's equations

Maxwell's equations connect the electric field **E** and the magnetic field intensity **H** to source charges ρ and currents **J** via the four relations

$$\boxed{\begin{array}{lcll} \nabla \cdot \mathbf{D} & = & \rho, & \text{Gauss's law} \\ \nabla \cdot \mathbf{B} & = & 0, & \text{Gauss's law} \\ \nabla \times \mathbf{E} & = & -\frac{\partial \mathbf{B}}{\partial t}, & \text{Faraday's law} \\ \nabla \times \mathbf{H} & = & \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}, & \text{Ampere's law.} \end{array}} \tag{16.1}$$

Here the source term ρ has units of charge per unit volume (C/m^3), and current source term **J** is in current per unit area A/m^2. **H** is related to the magnetic flux density **B** via $\mathbf{B} = \mu_0 \mathbf{H}$, and the displacement vector is related to the electric field via $\mathbf{D} = \epsilon_0 \mathbf{E}$. The constant ϵ_0 is the permittivity of vacuum, and μ_0 is the permeability of vacuum. They are related by $\epsilon_0 \mu_0 = 1/c^2$, where c is the speed of light in vacuum.

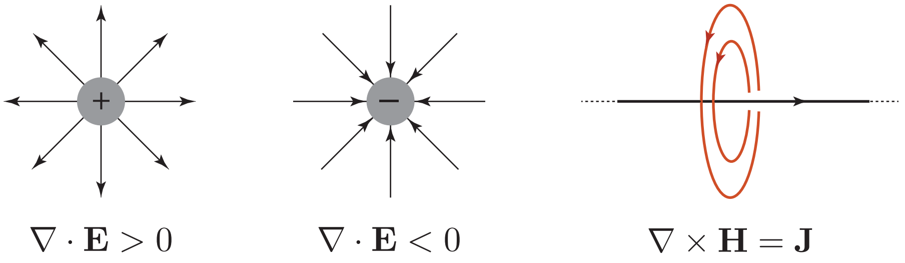

+
−
$\nabla \cdot \mathbf{E} > 0$
$\nabla \cdot \mathbf{E} < 0$
$\nabla \times \mathbf{H} = \mathbf{J}$

Fig. 16.2: Electrostatic Fields.

Gauss's law $\nabla \cdot \mathbf{E} = \rho/\epsilon_0$ says that electric field lines (vectors) due to *static* charges originate at points in space where there are +ve charges, and terminate at negative charges, as indicated in Figure 16.2. Vectors originating from a point in space have a positive divergence. This relation is also called the Poisson equation in semiconductor device physics, and if the charge is zero, it goes by the name of Laplace equation. Gauss's law for magnetic fields tells us that magnetic field lines **B** have no beginnings and no ends: unlike static electric field lines, they close on themselves.

Note that for electrostatics and magnetostatics, we put $\partial(...)/\partial t \to 0$, to obtain the static magnetic field relation $\nabla \times \mathbf{H} = \mathbf{J}$. The magnetic field lines curl around a wire carrying a dc current, as shown in Figure 16.2. Electrostatic phenomena such as electric fields in the presence of static charge such as p-n junctions, transistors, and optical devices in equilibrium, and magnetostatic phenomena such as magnetic fields near wires carrying dc currents are covered by the condition $\partial(...)/\partial t \to 0$, and electric and magnetic fields are *decoupled*. This means a static charge produces just electric fields and no magnetic fields. A static current (composed of charges moving at a *constant* velocity) produces a magnetic field, but no electric field.

Since in electrostatics, $\nabla \times \mathbf{E} = 0$, the static electric field vector can be expressed as the gradient of a scalar potential $\mathbf{E} = -\nabla\Phi$ because $\nabla \times (\nabla\Phi) = 0$ is an identity. Φ is then the scalar electric potential. However, the same cannot be done for the magnetic field vector even in static conditions, because $\nabla \times \mathbf{H} = \mathbf{J} \neq 0$. However, the magnetic field can be written as the *curl* of another vector field $\mathbf{B} = \nabla \times \mathbf{A}$, where **A** is called the magnetic *vector potential*. Hence from the Maxwell equations, $\mathbf{E} = -d\mathbf{A}/dt$.

Faraday's law says that a time-varying magnetic field creates an electric field. The electric field lines thus produced 'curl' around the magnetic field lines. Ampere's law says that a magnetic field intensity **H** may be produced not just by a conductor carrying current

$\mathbf{J}$, but also by a time-varying electric field in the form of the displacement current $\partial\mathbf{D}/\partial t$. The original Ampere's law did not have the displacement current. Maxwell realized that without it, the four constitutive equations would violate current continuity relations. To illustrate, without the displacement current term, $\nabla \times \mathbf{H} = \mathbf{J}$, and taking the divergence of both sides, we get $\nabla \cdot \nabla \times \mathbf{H} = \nabla \cdot \mathbf{J} = 0$ because the divergence of curl of any vector field is zero. But the continuity equation requires

$$\boxed{\nabla \cdot \mathbf{J} \quad = \quad -\partial\rho/\partial t, \quad \text{Continuity Equation}} \tag{16.2}$$

which is necessary for the conservation of charge. With the introduction of the displacement current term, Maxwell resolved this conflict: $\nabla \cdot \mathbf{J} = -\nabla \cdot \frac{\partial \mathbf{D}}{\partial t} = -\frac{\partial}{\partial t}(\nabla \cdot \mathbf{D}) = -\frac{\partial \rho}{\partial t}$, which connects to Gauss's law.

16.3 Light emerges from Maxwell's equations

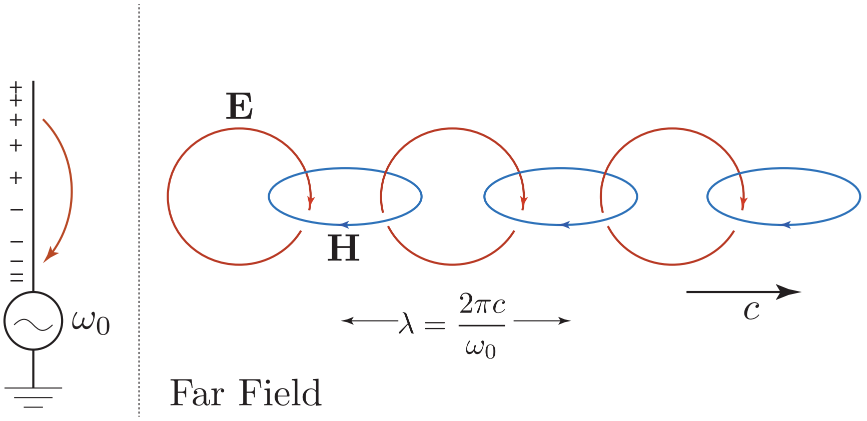


Fig. 16.3: Antenna producing an electromagnetic wave.

The displacement current term is the crucial link between electricity and magnetism, and leads to the existence of light as an electromagnetic wave. Let's first look at this feature qualitatively. Figure 16.3 shows a metal wire connected to an *ac* voltage source. The battery sloshes electrons back and forth from the ground into the wire, causing a charge-density wave as shown schematically. Note that the charge density in the wire is changing *continuously* in time and space. The frequency is ω_0. As a result of charge pileups, electric field lines emerge from +ve charges and terminate on -ve charges. This electric field is changing in space and time as well, leading to non-zero $\nabla \times \mathbf{E}$ and $\partial\mathbf{E}/\partial t$. The time-varying electric field *creates* a time-varying magnetic field $\mathbf{H}$ because of displacement current. The time-varying magnetic field creates a time-varying electric field by Faraday's law. Far from the antenna, the fields detach from the source antenna and become *self-sustaining*: the time-varying $\mathbf{E}$ creates $\mathbf{H}$, and vice versa. An electromagnetic wave is thus born; the oscillations of electric and magnetic fields move at the speed of light c. For an antenna radiating at a frequency ω_0, the wavelength is $\lambda = 2\pi c/\omega_0$. That the wave is self-sustaining is the most fascinating feature of light. If at some time the battery was switched off, the far field wave continues to propagate - *forever*, unless it encounters charges again. That of course is how light from the most distant galaxies and supernovae reach our antennas and telescopes, propagating through 'light years' in the vacuum of space, sustaining the oscillations[2].

[2] Boltzmann wrote "... was it a God who wrote these lines ..." in connection to "Let there be light".

Now let's make this observation mathematically rigorous. Consider a region in space with no charges ($\nabla \cdot \mathbf{D} = \rho = 0 = \nabla \cdot \mathbf{E}$) and no currents $\mathbf{J} = 0$. Take the curl of Faraday's equation to obtain $\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} = -\frac{\partial}{\partial t}(\nabla \times \mathbf{B}) = -\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{E}$, where we make use of Ampere's law. Since in a source-free region $\nabla \cdot \mathbf{E} = 0$, we get the *wave equations*

$$\boxed{\begin{array}{rcl} (\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2})\mathbf{E} & = & 0, \quad \text{Wave Equations} \\ (\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2})\mathbf{B} & = & 0. \end{array}} \tag{16.3}$$

Note that the wave equation states that the electric field and magnetic field oscillate both in space and time. The ratio of oscillations in space (captured by ∇^2) and oscillations in time (captured by $\frac{\partial^2}{\partial t^2}$) is the speed at which the wave moves, and it is $c = 1/\sqrt{\mu_0 \epsilon_0}$. The number is exactly equal to the experimentally measured speed of light, which solidifies the connection that light is an electromagnetic wave. We note that just like the solution to Dirac's equation in quantum mechanics is the electron, the solution of Maxwell's wave equation is light (or photons). Thus one can say that light has 'emerged' from the solution of Maxwell equations.

However, we must be cautious in calling the wave equation above representing light *alone*. Consider a generic wave equation $(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2})f(r,t) = 0$. This wave moves at a speed v. We can create a sound wave, *and* a water wave that moves at the same speed v, and $f(r,t)$ will represent distinct physical phenomena. If a cheetah runs as fast as a car, they are not the same object!

Consider a generic vector field of the type $\mathbf{V}(\mathbf{r},t) = V_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}\hat{\eta}$, where $\hat{\eta}$ is the direction of the vector. This field will satisfy the wave equations 16.3 if $\omega = c|\mathbf{k}|$, as may be verified by substitution. This requirement is the first constraint on the nature of electromagnetic waves. The second stringent constraint is that the field must satisfy Gauss's laws $\nabla \cdot \mathbf{E} = 0$ and $\nabla \cdot \mathbf{B} = 0$ for free space. In other words, electric and magnetic vector fields are a special class of vector fields. Their special nature is elevated by the physical observation that *no other wave can move at the speed of light.* Einstein's theory of relativity proves that the speed of light is absolute, and unique for electromagnetic waves: every other kind of wave falls short of the speed of light. Thus, Maxwell's wave equation *uniquely* represents light, self-sustaining oscillating electric and magnetic fields.

16.4 Maxwell's equations in $(\mathbf{k}, \omega)$ space

Consider an electromagnetic wave of a fixed frequency ω. Since $\mathbf{E}, \mathbf{B} \propto e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}$, we make two observations. Time derivatives of Faraday and Ampere's laws give $\frac{\partial}{\partial t}e^{-i\omega t} = -i\omega e^{-i\omega t}$, which means we can replace $\frac{\partial}{\partial t} \rightarrow -i\omega$, $\frac{\partial^2}{\partial t^2} \rightarrow (-i\omega)^2$, and so on. Similarly, the vector operators div and curl act on the $e^{i\mathbf{k}\cdot\mathbf{r}}$ part only, giving $\nabla \cdot (e^{i\mathbf{k}\cdot\mathbf{r}}\hat{\eta}) = i\mathbf{k} \cdot (e^{i\mathbf{k}\cdot\mathbf{r}}\hat{\eta})$ and $\nabla \times (e^{i\mathbf{k}\cdot\mathbf{r}}\hat{\eta}) = i\mathbf{k} \times (e^{i\mathbf{k}\cdot\mathbf{r}}\hat{\eta})$. These relations may be verified by straightforward substitution. Thus, we can replace $\nabla \rightarrow i\mathbf{k}$. With these observations, Maxwell equations in free-space become

$$\boxed{\begin{array}{rcl} \mathbf{k} \cdot \mathbf{E} & = & 0, \\ \mathbf{k} \cdot \mathbf{B} & = & 0, \\ \mathbf{k} \times \mathbf{E} & = & \omega \mathbf{B}, \\ \mathbf{k} \times \mathbf{B} & = & -\frac{\omega}{c^2}\mathbf{E}. \end{array}} \tag{16.4}$$

Note that we have converted Maxwell's equations in real space and time $(\mathbf{r}, t)$ to 'Fourier' space $(\mathbf{k}, \omega)$ in this process. Just as in Fourier analysis where we decompose a function into its spectral components, light of a particular $\mathbf{k}$ and corresponding frequency $\omega = c|\mathbf{k}|$ is spectrally pure, and forms the 'sine' and 'cosine' bases. Any mixture of light is a linear combination of these spectrally pure components: for example white light is composed of multiple wavelengths. Since $\mathbf{B} = \nabla \times \mathbf{A}$, we can write $\mathbf{B} = i\mathbf{k} \times \mathbf{A}$, and hence the magnitudes are related by $B^2 = k^2 A^2 = (\frac{\omega}{c})^2 A^2$. The energy content in a region in space of volume Ω that houses electric and magnetic fields of frequency ω is given by

$$\boxed{H_{em}(\omega) = \Omega \cdot [\frac{1}{2}\epsilon_0 E^2 + \frac{1}{2}\mu_0 H^2] = \Omega \cdot [\frac{1}{2}\epsilon_0 E^2 + \frac{1}{2}\epsilon_0 \omega^2 A^2]}. \tag{16.5}$$

If you have noticed a remarkable similarity between the expression for energy of an electromagnetic field with that of a harmonic oscillator (from Chapter **??**) $H_{osc} = \frac{\hat{p}^2}{2m} + \frac{1}{2}m\omega^2 x^2$, you are in luck. In Chapter **??**, this analogy will enable us to fully quantize the electromagnetic field, resulting in a rich new insights.

Let us now investigate the properties of a spectrally pure, or 'monochromatic' component of the electromagnetic wave. From equations 16.4, we note that $\mathbf{k} \perp \mathbf{E} \perp \mathbf{B}$, and the direction of $\mathbf{k}$ is along $\mathbf{E} \times \mathbf{B}$. The simplest possibility is shown in Figure 16.4. If we align the $x-$axis along the electric field vector and the $y-$axis along the magnetic field vector, then the wave propagates along the +ve $z-$axis, i.e., $\mathbf{k} = k\hat{z}$. The electric field vectors lie in the $x - z$ plane, and may be written as $\mathbf{E}(\mathbf{r}, t) = E_0 e^{i(kz - \omega t)}\hat{x}$, which is a plane wave. For a plane wave, nothing changes along the planes perpendicular to the direction of propagation, so the $\mathbf{E}$ field is the same at all $x - y$ planes: $\mathbf{E}(x, y, z) = \mathbf{E}(0, 0, z)$.

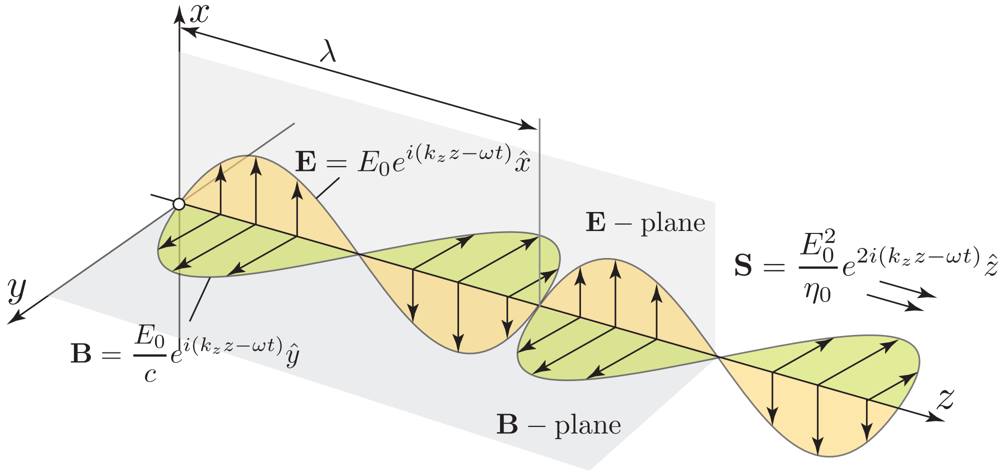

x
λ
$\mathbf{E} = E_0 e^{i(k_z z - \omega t)}\hat{x}$
$\mathbf{E}$ − plane
$\mathbf{S} = \frac{E_0^2}{\eta_0} e^{2i(k_z z - \omega t)}\hat{z}$
y
$\mathbf{B} = \frac{E_0}{c} e^{i(k_z z - \omega t)}\hat{y}$
$\mathbf{B}$ − plane
z

Fig. 16.4: Electromagnetic wave.

From Faraday's law, $\mathbf{B} = \mathbf{k} \times \mathbf{E}/\omega$, and the magnetic field vectors $\mathbf{B}(\mathbf{r}, t) = \frac{E_0}{c} e^{i(kz - \omega t)}\hat{y}$ lie in the $y - z$ plane. Note that here we use $\omega = ck$ and $k = k_z$. The amplitudes of the electric and magnetic fields are thus related by $E_0 = cB_0$, and the relation to magnetic field *intensity* $\mathbf{H} = \mathbf{B}/\mu_0$ is $E_0 = c\mu_0 H_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} H_0 = \eta_0 H_0$. Since E_0 has units V/m and H_0 has units A/m, η has units of V/A or Ohms. η_0 is called the impedance of free space; it has a value $\eta_0 \approx 377\Omega$.

The direction of propagation of this wave is *always* perpendicular to the electric and magnetic field vectors and given by the right hand rule. Since the field vectors lie on well-defined planes, this type of electromagnetic wave is called *plane-polarized*. In case there was a phase difference between the electric and magnetic fields, the electric and magnetic field vectors will rotate in the $x - y$ planes as the wave propagates, and the wave would then be called circularly or *elliptically* polarized, depending upon the phase difference.

For the monochromatic wave, Maxwell's wave equation becomes $(|\mathbf{k}|^2 - (\frac{\omega}{c})^2)\mathbf{E} = 0$. For non-zero $\mathbf{E}$, $\omega = c|\mathbf{k}| = ck$. The electromagnetic field carries energy in the +ve

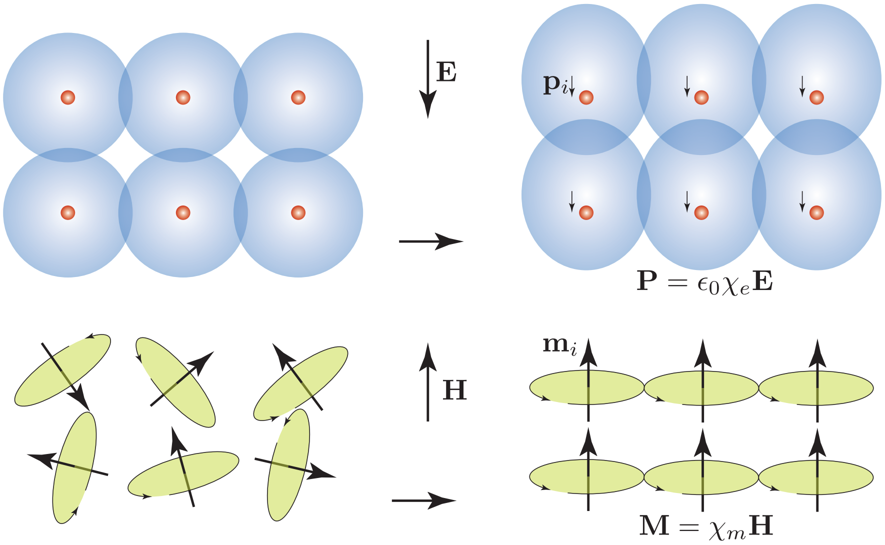
$z-$direction. The instantaneous power carried by the wave is given by the Poynting vector $\mathbf{S}(\mathbf{r},t) = \mathbf{E} \times \mathbf{H} = \frac{E_0^2}{\eta_0} e^{i(kz-\omega t)} \hat{z}$. The units are in Watts/m^2. Typically we are interested in the time-averaged power density, which is given by

$$\mathbf{S} = \langle \mathbf{S}(\mathbf{r},t) \rangle = \frac{1}{2} \mathrm{Re}[\mathbf{E} \times \mathbf{H}^{\star}] = \frac{E_0^2}{2\eta} \hat{z} = \frac{\eta}{2} H_0^2 \hat{z}, \tag{16.6}$$

where $\hat{z}$ is the direction of propagation of the wave. In later chapters, the energy carried by a monochromatic wave will for the starting point to understand the interaction of light with matter. In the next chapter, we will discuss how the energy carried by an electromagnetic wave as described by Equation 16.6 actually appears not in continuous quantities, but in quantum packets. Before we do that, we briefly discuss the classical picture of light interacting with material media.

16.5 Maxwell's equations in material media

How does light interact with a material medium? Running the video of the process of the *creation* of light in Figure 16.3 backwards, we can say that when an electromagnetic wave hits a metal wire, the electric field will slosh electrons in the wire back and forth generating an ac current. That is the principle of operation of a receiving antenna. What happens when the material does not have freely conducting electrons like a metal? For example, in a dielectric some electrons are tightly bound to atomic nuclei (core electrons), and others participate in forming chemical bonds with nearest neighbor atoms. The electric field of the electromagnetic wave will deform the electron clouds that are most 'flexible' and 'polarize' them. Before the external field was applied, the centroid of the negative charge from the electron clouds and the positive nuclei exactly coincided in space. When the electron cloud is deformed, the centroids do not coincide any more, and a net dipole is formed, as shown in Figure 16.5. The electric field of light primarily interacts with electrons that are most loosely bound and deformable; protons in the nucleus are far heavier, and held strongly in place in a solid medium. Let us give these qualitative observations a quantitative basis.



$\mathbf{E}$
$\mathbf{p}_i$
$\mathbf{P} = \epsilon_0 \chi_e \mathbf{E}$
$\mathbf{H}$
$\mathbf{m}_i$
$\mathbf{M} = \chi_m \mathbf{H}$

Fig. 16.5: Dielectric and Magnetic materials. Orientation of electric and magnetic dipoles by external fields, leading to electric and magnetic susceptibilities.

The displacement vector in free space is $\mathbf{D} = \epsilon_0 \mathbf{E}$. In the presence of a dielectric, it has an additional contribution $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$, where $\mathbf{P}$ is the polarization of the dielectric. The classical picture of polarization is an electric dipole $\mathbf{p}_i = q d_i \hat{n}$ in every unit cell of the solid. This dipole has zero magnitude in the absence of the external field[3]. The electric field of light stretches the electron cloud along it, forming dipoles along itself. Thus, $\mathbf{p}_i$ points along $\mathbf{E}$. The net polarization[4] is the volume-averaged dipole density $\mathbf{P} = \frac{1}{V}\sum_V \mathbf{p}_i$. Based on the material properties of the dielectric, we absorb all microscopic details into one parameter by writing

$$\mathbf{P} = \epsilon_0 \chi_e \mathbf{E}, \tag{16.7}$$

where the parameter χ_e is referred to as the electric susceptibility of the solid. With this definition, the displacement vector becomes

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \epsilon_0 \chi_e \mathbf{E} = \epsilon_0 \underbrace{(1+\chi_e)}_{\epsilon_r} \mathbf{E} = \epsilon \mathbf{E}, \tag{16.8}$$

where the dielectric property of the material is captured by the modified dielectric constant $\epsilon = \epsilon_0 \epsilon_r = \epsilon_0(1+\chi_e)$. The *relative* dielectric constant is 1 plus the electric susceptibility of the material. Clearly the relative dielectric constant of vacuum is 1 since there are no atoms to polarize and nothing is 'susceptible'.

In exactly the same way, if the material is magnetically polarizable, then $\mathbf{B} = \mu_0(\mathbf{H}+\mathbf{M})$, where $\mathbf{M}$ is the magnetization vector. If there are tiny magnetic dipoles $\mathbf{m}_i = IA\hat{n}$ formed by circular loops carrying current I in area A in the material medium (see Figure 16.5), the macroscopic magnetization is given by $\mathbf{M} = \frac{1}{V}\sum_V \mathbf{m}_i = \chi_m \mathbf{H}$, which leads to the relation

$$\mathbf{B} = \mu_0(\mathbf{H} + \chi_m \mathbf{H}) = \mu_0 \underbrace{(1+\chi_m)}_{\mu_r} \mathbf{H} = \mu \mathbf{H}, \tag{16.9}$$

With these changes, the original Maxwell equations remain the same, but now $\mathbf{D} = \epsilon \mathbf{E}$ and $\mathbf{B} = \mu \mathbf{H}$, so we make the corresponding changes $\epsilon_0 \to \epsilon = \epsilon_0 \epsilon_r$ and $\mu_0 \to \mu = \mu_0 \mu_r$ everywhere. For example, the speed of light in a material medium then becomes $v = \frac{1}{\sqrt{\mu\epsilon}} = \frac{c}{\sqrt{\epsilon_r \mu_r}}$. If the material is non-magnetic, then $\mu_r = 1$, and $v = \frac{c}{\sqrt{\epsilon_r}} = \frac{c}{n}$, where $n = \sqrt{\epsilon_r}$ is called the refractive index of the material. Thus light travels slower in a material medium than in free space. Similarly, the wave impedance becomes $\eta_0 \to \eta = \sqrt{\frac{\mu}{\epsilon}} = \frac{\eta_0}{n}$ where the right equality holds for a non-magnetic medium.

If the material medium is conductive, or can absorb the light through electronic transitions, then the phenomena of *absorption* and corresponding attenuation of the light is captured by introducing an imaginary component to the dielectric constant, $\epsilon \to \epsilon_R + i\epsilon_I$. This leads to an imaginary component of the propagation vector $\mathbf{k}$, which leads to attenuation. We will see in Chapters **??** and **??** how we can calculate the absorption coefficients from quantum mechanics.

Electric and magnetic field lines may cross interfaces of different material media. Then, the Maxwell equations provide rules for tracking the magnitudes of the tangential and perpendicular components. These **boundary conditions** are given by

$$\boxed{\begin{array}{lcl} \mathbf{E}_{1t} - \mathbf{E}_{2t} & = & 0, \\ \mathbf{H}_{1t} - \mathbf{H}_{2t} & = & \mathbf{J}_s \times \hat{\mathbf{n}}, \\ D_{1n} - D_{2n} & = & \rho_s, \\ B_{1n} - B_{2n} & = & 0. \end{array}} \tag{16.10}$$

[3] Except in materials that have spontaneous, piezoelectric, or ferroelectric polarization.

[4] This classical picture of polarization is not consistent with quantum mechanics. The quantum theory of polarization requires the concept of Berry phases, which is the subject of Chapter **??**.

In words, the boundary condition relations say that the tangential component of the electric field $\mathbf{E}_t$ is *always* continuous across an interface, but the normal component is discontinuous if there are charges at the interface. If there are no *free* charges at the interface ($\rho_s = 0$), $\epsilon_1 E_{1n} = \epsilon_2 E_{2n}$, implying the normal component of the electric field is larger in the material with a smaller dielectric constant. This feature is used in Si MOSFETs, where much of the electric field drops across an oxide layer rather than in the semiconductor which has a higher dielectric constant. Similarly, the normal component of the magnetic field is always continuous across an interface, whereas the tangential component can change if there is a surface current flowing at the interface of the two media.

The force in Newtons on a particle of charge q in the presence of an electric and magnetic field is given by the Lorentz equation

$$\boxed{\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}).} \tag{16.11}$$

Since the energy of the charged particle changes as $W = \int \mathbf{F} \cdot d\mathbf{r}$, the rate of change of energy is $\mathbf{F} \cdot \mathbf{v} = q\mathbf{E} \cdot \mathbf{v}$, which is the power delivered to the charged particle by the fields. Note that a static magnetic field cannot deliver power since $\mathbf{v} \times \mathbf{B} \cdot \mathbf{v} = 0$. Thus a time-independent magnetic field cannot change the energy of a charged particle. But a time-dependent magnetic field creates an electric field, which can.

When a point charge is *accelerated* with acceleration a, it radiates electromagnetic waves. Radiation travels at the speed of light. So the electric and magnetic fields at a point far from the charge are determined by a *retarded* response. Using retarded potentials, or more intuitive approaches[5], one obtains that the radiated electric field goes as

$$\mathbf{E}_r = (\frac{qa}{4\pi\epsilon_0 c^2})\frac{\sin\theta}{r}\hat{\theta}, \tag{16.12}$$

expressed in spherical coordinates with the charge at the origin, and accelerating along the $x-$axis. The radiated magnetic field $\mathbf{H}_r$ curls in the $\hat{\phi}$ direction and has a magnitude $|\mathbf{E}_r|/\eta_0$. The radiated power is obtained by the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$ as

$$\mathbf{S} = (\frac{\mu_0 q^2 a^2}{16\pi^2 c^2})(\frac{\sin\theta}{r})^2 \hat{r}, \tag{16.13}$$

Note that unlike static charges or currents that fall as $1/r^2$ away from the source, the *radiated* $\mathbf{E}$ and $\mathbf{H}$ fields fall as $1/r$. If they didn't, the net power radiated very far from the source will go to zero since $\oint \mathbf{S} \cdot d\mathbf{A} \sim S(r) 4\pi r^2 \to 0$. Integrating the power over the angular coordinates results in the famous **Larmor Formula** for the net electromagnetic power in Watts radiated by an accelerating charge:

$$\boxed{P = \frac{\mu_0 q^2 a^2}{6\pi c}} \tag{16.14}$$

16.6 Need for a quantum theory of light

Classical electromagnetism contained in Maxwell's equations can explain a remarkably large number of experimentally observed phenomena, but not all. We discussed in the beginning of this chapter that radiation of electromagnetic waves can be created in an antenna, which in its most simple form is a conducting wire in which electrons are sloshed back and forth. The *collective* acceleration, coupled with the Larmor formula can explain radiation from a vast number of sources of electromagnetic radiation.

By the turn of the 20th century, improvements in spectroscopic equipment had helped resolve what was originally thought as broadband (many frequencies ω) radiation into the

[5] An intuitive picture for radiation by an accelerating charge was first given by J. J. Thomson, the discoverer of the electron.

purest spectral components. It was observed that different gases had different spectral signatures. The most famous among them were the spectral features of the hydrogen atom, then known as the hydrogen gas. There is nothing *collective* about hydrogen gas, since it is not a conductor and there are not much electrons to slosh around as a metal. The classical theory for radiation proved difficult to apply to explain the spectral features. Classical electromagnetism could not explain the photoelectric effect, and the spectrum of blackbody radiation either. The search for an explanation led to the quantum theory of light, which is the subject of the next chapter.

ECE 4070, Spring 2017
Physics of Semiconductors and Nanostructures
Handout 17

Light-Matter Interaction

Contents

17.1 Introduction 165
17.2 Electron-photon matrix elements for semiconductors 165
17.3 The absorption spectrum of bulk semiconductors 167
17.4 The number of photons in light 168
17.5 Photon absorption rate in bulk semiconductors 169
17.6 The Equilibrium Absorption Coefficient $\alpha_0(\hbar\omega)$ 174

17.1 Introduction

In this chapter, we explore fundamental optical transitions in bulk 3-dimensional semiconductors. We approach the topic by first investigating the optical absorption spectrum. The spectrum will direct us to a rich range of electron state transitions affected by the electron-photon interaction. Then, we explore the most important of these transitions: interband (valence → conduction) transitions in more detail. We derive expressions for the equilibrium interband absorption coefficient $\alpha_0(\hbar\omega)$ for bulk semiconductors. With the understanding of the physics of optical absorption, in the next chapter we extend the concept to non-equilibrium situations to explain optical emission, optical gain, inversion, and lasing conditions. The key to understanding these concepts is a clear quantum-mechanical picture of optical transitions, and the role of non-equilibrium conditions. We begin with the fundamental quantum-mechanical optical transitions by recalling the electron-photon Hamiltonian.

17.2 Electron-photon matrix elements for semiconductors

The Hamiltonian with the magnetic vector potential $\mathbf{A}$ in the form

$$\boxed{H = \frac{1}{2m_0}(\hat{\mathbf{p}} + e\mathbf{A})^2 + V(\mathbf{r})} \tag{17.1}$$

captures the interaction of electrons with light. This form of the Hamiltonian explains the interaction of light with atoms, and successfully explains the optical spectra of atoms. The electron energy spectra of atoms are typically very sharp because of the discrete energy eigenvalues of electrons. Here, we apply the same idea to bulk semiconductors, in which the energy eigenvalues form bands separated by energy gaps.

We recall that the electromagnetic wave enters the Hamiltonian via the magnetic vector potential $\mathbf{A}$, which is related to the electric field via

$$\nabla \times \mathbf{E}(\mathbf{r},t) = -\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r},t) \underbrace{\rightarrow}_{\mathbf{B}(\mathbf{r},t)=\nabla\times\mathbf{A}(\mathbf{r},t)} \mathbf{E}(\mathbf{r},t) = -\frac{\partial}{\partial t}\mathbf{A}(\mathbf{r},t), \tag{17.2}$$

and we work in the Coulomb gauge

$$\nabla \cdot \mathbf{A} = 0. \tag{17.3}$$

This enables the vector potential $\mathbf{A}$ to commute with the momentum operator $\hat{\mathbf{p}}$

$$[\hat{\mathbf{p}}, \mathbf{A}] = 0 \to \hat{\mathbf{p}} \cdot \mathbf{A} = \mathbf{A} \cdot \hat{\mathbf{p}}, \tag{17.4}$$

which leads to the electron-photon Hamiltonian

$$H = \underbrace{[\frac{\hat{p}^2}{2m_0} + V(\mathbf{r})]}_{\hat{H}_0} + \underbrace{\frac{e}{m_0}\mathbf{A} \cdot \hat{\mathbf{p}}}_{W} + \underbrace{\frac{e^2A^2}{2m_0}}_{neglect}. \tag{17.5}$$

We have written out the Hamiltonian in terms of the electron Hamiltonian $\hat{H}_0$, and the 'perturbation' term seen by the electron due to the electromagnetic wave. For an electron in a semiconductor crystal, the potential energy term in the unperturbed Hamiltonian is the periodic crystal potential $V(\mathbf{r} + \mathbf{a_0}) = V(\mathbf{r})$, where $\mathbf{a_0}$ is a lattice constant. We neglect the perturbation term that goes as the square of the magnetic vector potential for 'weak' intensities of light. This is justified when the condition $|e\mathbf{A}| << |\mathbf{p}| \sim \hbar\pi/a_0$ is met; in other words, we neglect the term $\frac{e^2A^2}{2m_0}$ w.r.t. $\frac{\hat{p}^2}{2m_0}$. The net Hamiltonian we retain then has the electron experiencing a perturbation

$$\boxed{\hat{W} = \frac{e}{m_0}\mathbf{A} \cdot \hat{\mathbf{p}}} \tag{17.6}$$

due to its interaction with light. The magnetic vector potential for an EMag wave is of the form[1]

$$\mathbf{A}(\mathbf{r}, \mathbf{t}) = \hat{e}A_0 \cos(\mathbf{k_{op}} \cdot \mathbf{r} - \omega t) \tag{17.7}$$

$$= \hat{e}\frac{A_0}{2}e^{+i\mathbf{k_{op}}\cdot\mathbf{r}}e^{-i\omega t} + \hat{e}\frac{A_0}{2}e^{-i\mathbf{k_{op}}\cdot\mathbf{r}}e^{+i\omega t}, \tag{17.8}$$

where ω is the angular frequency of the EMag wave, $\hat{e}$ is the unit vector along the electric (and vector potential) field, and $\mathbf{k_{op}}$ is the propagation wave vector of magnitude $2\pi/\lambda$. The electron-photon interaction Hamiltonian is then given by

$$\hat{W}(\mathbf{r}, t) = \frac{e}{m_0}\mathbf{A} \cdot \hat{\mathbf{p}} \tag{17.9}$$

$$= \hat{W}(\mathbf{r})e^{-i\omega t} + \hat{W}^{+}(\mathbf{r})e^{+i\omega t} \tag{17.10}$$

$$\boxed{\hat{W}(\mathbf{r}) = \frac{eA_0e^{i\mathbf{k_{op}}\cdot\mathbf{r}}}{2m_0}\hat{e} \cdot \hat{\mathbf{p}}} \tag{17.11}$$

$$\boxed{\hat{W}^{+}(\mathbf{r}) = \frac{eA_0e^{-i\mathbf{k_{op}}\cdot\mathbf{r}}}{2m_0}\hat{e} \cdot \hat{\mathbf{p}}} \tag{17.12}$$

The electron-photon matrix elements for bulk semiconductors are thus of the form $\langle\mathbf{k_c}|\hat{W}|\mathbf{k_v}\rangle$ and $\langle\mathbf{k_c}|\hat{W}^{+}|\mathbf{k_v}\rangle$, where the *unperturbed* electron states $|\mathbf{k_c}\rangle$ and $|\mathbf{k_v}\rangle$ are solutions of the unperturbed Hamiltonian $\hat{H}_0 = \frac{\hat{p}^2}{2m_0} + V(\mathbf{r})$. But this is precisely what we discussed in chapters **??** and **??** for semiconductors. The electron states in the valence and conduction bands in the effective mass approximation are $\psi_c(\mathbf{r}) = \langle\mathbf{r}|\mathbf{k_c}\rangle = (\frac{e^{i\mathbf{k_c}\cdot\mathbf{r}}}{\sqrt{V}})u_c(\mathbf{r})$ for bulk semiconductors. The term in the round bracket is a slowly varying envelope function,

[1] This approach of treating the electron-photon interaction is *semi-classical*, justified for *classical* electromagnetic fields when the number of photons is much larger than unity. It is semi-classical because electrons receive the full quantum treatment, but we neglect the quantization of the electromagnetic field, treating it as a classical infinite source or sink of energy, albeit in quantum packets of $\hbar\omega$. The electromagnetic field will be quantized in Chapters **??** and beyond in this book.

and $u_c(\mathbf{r})$ is the periodic part of the Bloch function. The effective mass approximation transforms the unperturbed electronic Hamiltonian into the much simpler form $\frac{\hat{p}^2}{2m_0} + V(\mathbf{r}) \rightarrow \frac{\hat{p}^2}{2m_c^\star}$, and the corresponding effective-mass Schrodinger equation is $\frac{\hat{p}^2}{2m^\star}\psi_c(\mathbf{r}) = (E - E_c)\psi_c(\mathbf{r})$. We will work in this effective-mass theory. The advantage of working in the effective-mass theory is that the light-matter interaction matrix elements for electrons confined in low-dimensional structures such as quantum wells, wires, or dots follows in a simple way from the bulk results. We will need the matrix elements shortly to explain the absorption spectra of bulk semiconductors, which we discuss next.

17.3 The absorption spectrum of bulk semiconductors

We learn early of the Beer-Lambert 'law', which states that if light of intensity I_0 is incident on a material that absorbs, the intensity will decay inside the material as $I(z) = I_0 e^{-\alpha z}$. Here α is the absorption coefficient, in units of inverse length. Typically the unit used for α is cm^{-1}. Let us consider the following experiment: take a piece of bulk semiconductor, say GaAs or GaN, and using a tunable light source, measure α as a function of the photon energy $\hbar\omega$. Then we obtain the absorption spectrum $\alpha(\hbar\omega)$. The absorption spectrum of most semiconductors looks like what is shown in the schematic Figure 17.1.

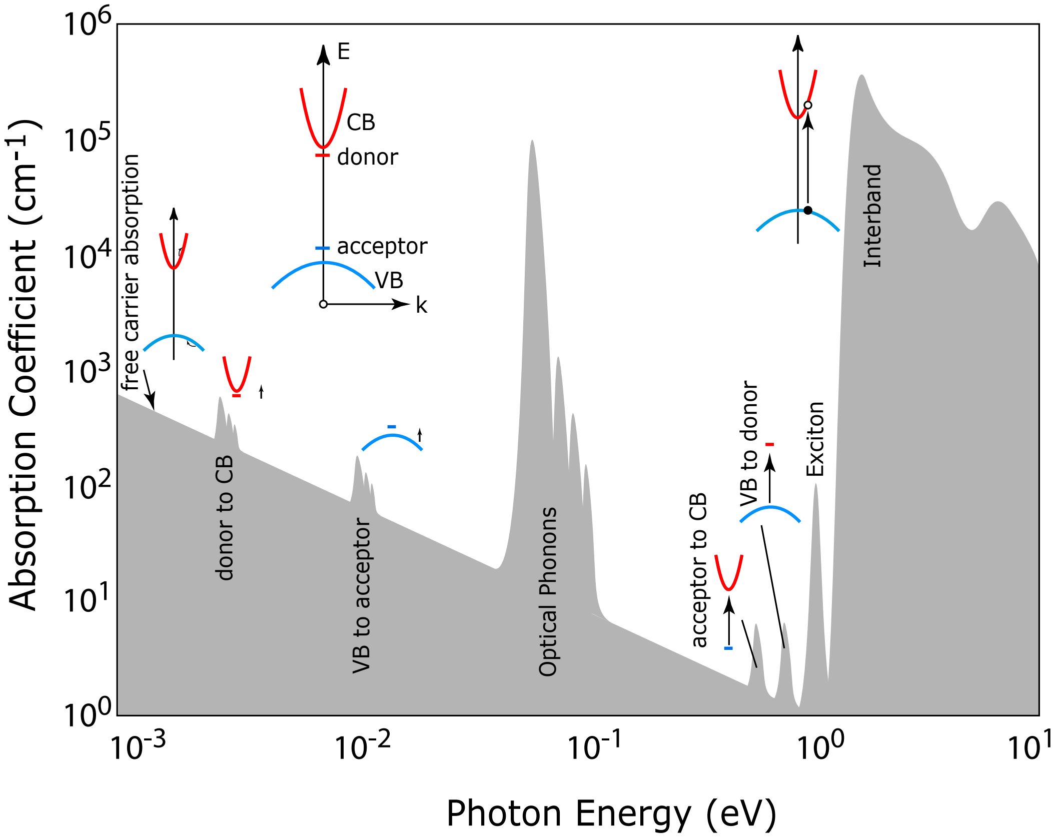


Fig. 17.1: Schematic absorption spectrum $\alpha(\hbar\omega)$ of bulk semiconductors. The insets depict various state transitions upon absorption of photons.

The inset of Figure 17.1 indicates the electron bandstructure of the bulk semiconductor, including states corresponding to donor and acceptor dopants. The transitions between electron states caused by photon absorption are indicated. The floor of the absorption spectrum is due to intraband transitions caused by the absorption of low energy ($\sim$ few meV) photons by free carriers. Transitions between dopant and band states are shown, in addition to the below-bandgap excitonic transition. Such optical measurements provide a sensitive experimental determination of dopant and excitonic energies with respect to the fundamental band edge energies. Photons can excite *mechanical* vibrations of the bulk semiconductor crystal by creating optical phonons: the absorption peak for this process is rather strong, and forms the basis or Raman spectroscopy. By far, the strongest absorption occurs for interband transitions, which is the focus of this chapter.

The absorption spectrum is quantitatively defined as

$$\boxed{\alpha(\hbar\omega) = \frac{\text{Number of photons absorbed per unit volume per second}}{\text{Number of photons incident per unit area per second}} = \frac{R(\hbar\omega)}{N_{ph}(\hbar\omega)}}. \tag{17.13}$$

In the next section we derive an expression for the denominator $N_{ph}(\hbar\omega)$, and in the following section we deal with the numerator $R(\hbar\omega)$.

17.4 The number of photons in light

Consider a monochromatic EMag wave of frequency ω and corresponding wavevector $\mathbf{k_{op}} = \frac{2\pi}{\lambda}\hat{n}$. For a plane wave, the magnetic vector potential is

$$\mathbf{A}(\mathbf{r}, \mathbf{t}) = \hat{e}A_0 \cos(\mathbf{k_{op}} \cdot \mathbf{r} - \omega t), \tag{17.14}$$

from where the electric field is obtained by using

$$\mathbf{E}(\mathbf{r}, t) = -\frac{\partial}{\partial t}\mathbf{A}(\mathbf{r}, t) \tag{17.15}$$

$$= -\hat{e}\omega A_0 \sin(\mathbf{k_{op}} \cdot \mathbf{r} - \omega t), \tag{17.16}$$

and the magnetic field intensity is

$$\mathbf{H}(\mathbf{r}, t) = \frac{1}{\mu}\nabla \times \mathbf{A}(\mathbf{r}, t) \tag{17.17}$$

$$= -\frac{1}{\mu}\mathbf{k_{op}} \times \hat{e}A_0 \sin(\mathbf{k_{op}} \cdot \mathbf{r} - \omega t). \tag{17.18}$$

Here we have used $\nabla \times (...) \equiv -\mathbf{k_{op}} \times (...)$ for plane waves, as described in Chapter **??**. Then, the energy carried by the plane wave per unit area per unit time is given by the Poynting vector

$$\mathbf{S}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) \times \mathbf{H}(\mathbf{r}, t) \tag{17.19}$$

$$= \mathbf{k_{op}}\frac{\omega A_0^2}{\mu}\sin^2(\mathbf{k_{op}} \cdot \mathbf{r} - \omega t) \tag{17.20}$$

Where we use the identity $\hat{e} \times \mathbf{k_{op}} \times \hat{e} = \mathbf{k_{op}}$. Since the frequency of typical UV-visible-IR light is very high, we time-average the Poynting vector over a period to obtain

$$\langle \mathbf{S}(\mathbf{r}, t) \rangle = \frac{\omega A_0^2}{2\mu}\mathbf{k_{op}}, \tag{17.21}$$

and its magnitude is

$$S = |\langle \mathbf{S}(\mathbf{r},t)\rangle| = \frac{\omega A_0^2}{2\mu} k_{op} = \frac{n_r c \epsilon_0 \omega^2 A_0^2}{2} = \frac{E_0^2}{2\eta} \tag{17.22}$$

where $\mu = \mu_0$ and $n_r = \sqrt{\mu_r \epsilon_r}$ is the refractive index of the media, in which the speed of light is c/n_r. Also note that $E_0 = \omega A_0$, and $\eta = \sqrt{\mu/\epsilon}$ is the field impedance. This relation gives us a way to find the magnitude of the vector potential A_0 if we know the power carried per unit area by the electromagnetic wave. Since energy in electromagnetic waves is carried in quantum packets (photons) of individual energy $\hbar\omega$, the *number of photons* that cross unit area per unit time is then given by

$$\boxed{N_{ph}(\hbar\omega) = \frac{S}{\hbar\omega} = \frac{n_r c \epsilon_0 \omega^2 A_0^2}{2\hbar\omega} = \frac{E_0^2}{2\eta\hbar\omega}.} \tag{17.23}$$

The *intensity* of light is proportional to the *square* of the electric (or magnetic) field amplitude, and thus the number of photons is a measure of the intensity of radiation. Equation 17.23 provides the denominator of the expression for absorption coefficient Equation 17.13. The numerator term is discussed in the next section.

17.5 Photon absorption rate in bulk semiconductors

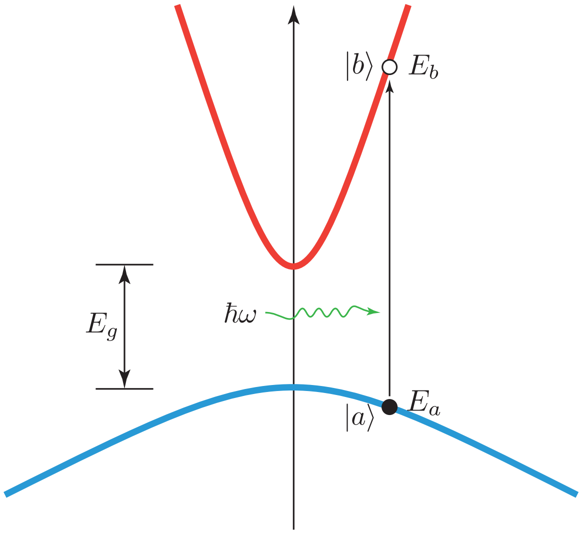


Fig. 17.2: The absorption process of a single photon by interband transition.

To find the rate of photon absorption in the bulk semiconductor, we apply Fermi's golden rule derived in Chapter 14. We first note that the numerator of Equation 17.13 has units of number of photons absorbed per unit volume per second. Consider Figure 17.2. An electron in the valence band state $|a\rangle$ absorbs a photon of energy $\hbar\omega$ and transitions into state $|a\rangle$ in the conduction band. Each such transition results in the *annihilation* of a photon from the EMag field. The rate at which this happens is given by Fermi's golden rule as

$$\frac{1}{\tau_{a\to b}} = \frac{2\pi}{\hbar}|\langle b|W(\mathbf{r})|a\rangle|^2\delta[E_b - (E_a + \hbar\omega)], \tag{17.24}$$

where $\langle b|W(\mathbf{r})|a\rangle$ is the perturbation matrix element, and the Dirac-delta function is a statement of energy conservation in the process. The reverse process of photon emission is also allowed, which results in the *creation* of a photon in the EMag field at the rate

$$\frac{1}{\tau_{b\to a}} = \frac{2\pi}{\hbar}|\langle a|W(\mathbf{r})|b\rangle|^2\delta[E_a - (E_b - \hbar\omega)], \tag{17.25}$$

which must be *subtracted* because an emission process makes a negative contribution to the number of photons absorbed. The above results are for the single states $|a\rangle$ and $|b\rangle$. A semiconductor crystal has a large number of states in the respective bands, so let's sum the rates for all possible transitions, and divide it by the net volume V to obtain the absorption rate per unit volume (in s$^{-1}\cdot$cm^{-3}). Add in the electron spin degeneracy $g_s = 2$ for each $\mathbf{k}$ state[2]. For the absorption process to occur, the lower state $|a\rangle$ has to be occupied (probability $= f_a$) and the higher state $|b\rangle$ has to be empty (probability $= (1 - f_b)$), where f's are the occupation functions. The net absorption rate per unit volume is then given by

$$R_{abs} = \frac{2}{V}\sum_{\mathbf{k_a}}\sum_{\mathbf{k_b}}\frac{2\pi}{\hbar}|W_{ba}|^2\delta[E_b - (E_a + \hbar\omega)]f_a(1 - f_b), \tag{17.26}$$

and the net emission rate per unit volume is

$$R_{em} = \frac{2}{V}\sum_{\mathbf{k_a}}\sum_{\mathbf{k_b}}\frac{2\pi}{\hbar}|W_{ab}|^2\delta[E_a - (E_b - \hbar\omega)]f_b(1 - f_a). \tag{17.27}$$

The summation runs over *all* valence band electron states $\mathbf{k_a}$ and conduction band electron states $\mathbf{k_b}$, including those that do not meet the criteria $E_b - E_a = \hbar\omega$. The energy conservation requirement is automatically taken care of by the Dirac-delta functions. We note now that the Dirac-delta functions are the same for emission and absorption process because $\delta[+x] = \delta[+x]$, $|W_{ab}| = |W_{ba}|$, and $f_a(1 - f_b) - f_b(1 - f_a) = f_a - f_b$. Therefore, the net photon *absorption* rate per unit volume is the difference

$$\boxed{R(\hbar\omega) = R_{abs} - R_{em} = \frac{2}{V}\sum_{\mathbf{k_a}}\sum_{\mathbf{k_b}}\frac{2\pi}{\hbar}|W_{ab}|^2\delta[E_b - (E_a + \hbar\omega)] \times (f_a - f_b)} \tag{17.28}$$

To evaluate the sum over states, we must first obtain an expression for the matrix element, which is given by the electron-photon perturbation term

$$W_{ab} = \langle b|\frac{e}{m_0}\mathbf{A}\cdot\hat{\mathbf{p}}|a\rangle. \tag{17.29}$$

At this stage, we need to know the wavefunctions corresponding to the band states $|a\rangle$ and $b\rangle$. In the effective mass approximation, the electron wavefunction = (envelope function) × (Bloch function). The valence band state wavefunction is then

$$\psi_a(\mathbf{r}) = C(\mathbf{r})u_v(\mathbf{r}) = \underbrace{\frac{e^{i\mathbf{k_v}\cdot\mathbf{r}}}{\sqrt{V}}}_{\text{Envelope } C(\mathbf{r})}\underbrace{u_v(\mathbf{r})}_{\text{Bloch}}, \tag{17.30}$$

[2] Photons carry an angular momentum of ±ℏ depending upon their polarization. Therefore, the conservation of angular momentum couples specific spin states. Here we are considering light with photons of mixed polarization. Anglular momentum conservation dictates which bands can be involved in the absorption or emission process, thus providing a way to selectively excite say the light hole, heavy hole, or split-off bands because they differ in their net angular momentum.

and the conduction band state wavefunction is

$$\psi_b(\mathbf{r}) = C'(\mathbf{r})u_c(\mathbf{r}) = \frac{e^{i\mathbf{k_c}\cdot\mathbf{r}}}{\sqrt{V}}u_c(\mathbf{r}). \tag{17.31}$$

Since the spatial part of the vector potential for the EMag wave is $\mathbf{A} = \hat{e}\frac{A_0}{2}e^{i\mathbf{k_{op}}\cdot\mathbf{r}}$, we obtain the matrix element $W_{ab} = \langle b|\frac{e}{m_0}\mathbf{A}\cdot\hat{\mathbf{p}}|a\rangle$ to be

$$W_{ab} = \frac{eA_0}{2m_0}\hat{e}\cdot(\int \psi_b^\star e^{i\mathbf{k_{op}}\cdot\mathbf{r}}\hat{\mathbf{p}}\psi_a d^3\mathbf{r}) \tag{17.32}$$

$$= \frac{eA_0}{2m_0}\hat{e}\cdot\int[\frac{e^{i\mathbf{k_c}\cdot\mathbf{r}}}{\sqrt{V}}u_c(\mathbf{r})]^\star \underbrace{(e^{i\mathbf{k_{op}}\cdot\mathbf{r}}\hat{\mathbf{p}})}_{\text{operator}}[\frac{e^{i\mathbf{k_v}\cdot\mathbf{r}}}{\sqrt{V}}u_v(\mathbf{r})]d^3\mathbf{r} \tag{17.33}$$

$$= \frac{eA_0}{2m_0}\hat{e}\cdot\int[e^{-i\mathbf{k_c}\cdot\mathbf{r}}u_c^\star(\mathbf{r})]\underbrace{(e^{i\mathbf{k_{op}}\cdot\mathbf{r}}\mathbf{p})}_{\text{operator}}[e^{+i\mathbf{k_v}\cdot\mathbf{r}}u_v(\mathbf{r})]\frac{d^3\mathbf{r}}{V} \tag{17.34}$$

$$= \frac{eA_0}{2m_0}\hat{e}\cdot\int[e^{-i\mathbf{k_c}\cdot\mathbf{r}}u_c^\star(\mathbf{r})](e^{i\mathbf{k_{op}}\cdot\mathbf{r}})[e^{+i\mathbf{k_v}\cdot\mathbf{r}}(\hbar\mathbf{k_v}u_v(\mathbf{r}) - i\hbar\nabla u_v(\mathbf{r}))]\frac{d^3\mathbf{r}}{V} \tag{17.35}$$

$$= \underbrace{\frac{eA_0}{2m_0}\hat{e}\cdot\int e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{r}}[u_c^\star(\mathbf{r})u_v(\mathbf{r})](\hbar\mathbf{k_v})\frac{d^3\mathbf{r}}{V}}_{\text{forbidden}} + \tag{17.36}$$

$$\underbrace{\frac{eA_0}{2m_0}\hat{e}\cdot\int e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{r}}[u_c^\star(\mathbf{r})\hat{\mathbf{p}}u_v(\mathbf{r})]\frac{d^3\mathbf{r}}{V}}_{\text{allowed}} \tag{17.37}$$

The first term is labeled *forbidden* because the integral is $\approx \hbar\mathbf{k_v}\langle\mathbf{k_c}|\mathbf{k_v}\rangle = 0$ if we neglect the photon momentum. This is because the states belong to different bands, and are orthogonal. The 'allowed' transition matrix element is:

$$W_{ab} = \frac{eA_0}{2m_0}\hat{e}\cdot\int_V e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{r}}[u_c^\star(\mathbf{r})\hat{\mathbf{p}}u_v(\mathbf{r})]\frac{d^3\mathbf{r}}{V} \tag{17.38}$$

$$= \frac{eA_0}{2m_0}\hat{e}\cdot\int_V \underbrace{e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{r}}}_{\text{slow}}\underbrace{[u_c^\star(\mathbf{r})\hat{\mathbf{p}}u_v(\mathbf{r})]}_{\text{periodic}}\frac{d^3\mathbf{r}}{\underbrace{N\Omega}_{V}} \tag{17.39}$$

$$= \frac{eA_0}{2m_0}\hat{e}\cdot\int_V \underbrace{\frac{e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{r}}}{N}}_{\text{slow}}\underbrace{[u_c^\star(\mathbf{r})\hat{\mathbf{p}}u_v(\mathbf{r})]}_{\text{periodic}}\frac{d^3\mathbf{r}}{\Omega} \tag{17.40}$$

To visualize the slow and periodic parts inside the integral, refer to Figure 17.3. The periodic term functions $u_c(\mathbf{r})\hat{\mathbf{p}}\mathbf{u_c}(\mathbf{r})$ repeats in every unit cell in real space of volume Ω of the crystal. But the slowly varying function of the form $e^{i\mathbf{k}\cdot\mathbf{r}}$ hardly changes inside a unit cell, it changes appreciably only over many many cells. Then, we treat the slowly varying function as constant inside a the unit cell located at $\mathbf{R}_i$, but the value to change from cell to cell. Then, the integral decomposes to

$$W_{ab} = \frac{eA_0}{2m_0}\hat{e}\cdot\underbrace{[\frac{\sum_{n=1}^{N}e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{R_n}}}{N}]}_{\delta_{\mathbf{k_c},\mathbf{k_v}+\mathbf{k_{op}}}}\underbrace{\int_\Omega[u_c^\star(\mathbf{r})\hat{\mathbf{p}}u_v(\mathbf{r})]\frac{d^3\mathbf{r}}{\Omega}}_{\mathbf{p_{cv}}}. \tag{17.41}$$

The sum runs over all unit cells in real space. Since $\sum_{n=1}^{N}e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{R_n}}$ is the sum of the complex exponential at every unit cell site $\mathbf{R}_n$, and there are a *lot* of them, let us visualize this sum. Refer to Figure 17.4 to see why the sum $\frac{\sum_{n=1}^{N}e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{R_n}}}{N}$ is zero

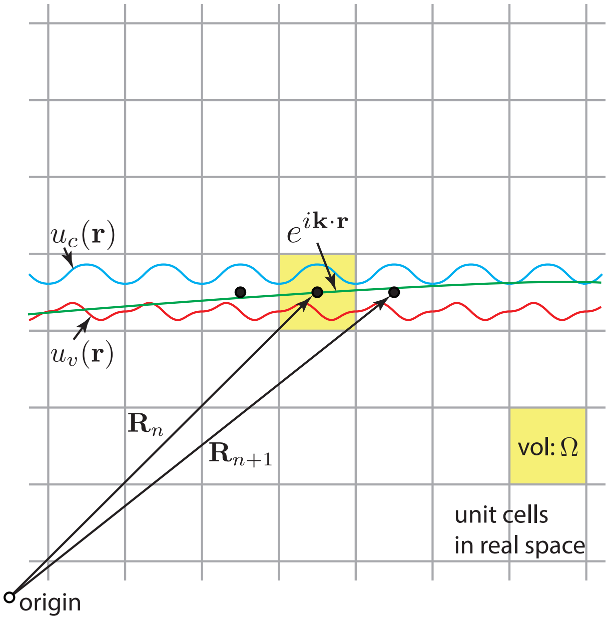


Fig. 17.3: Explanation of the decomposition of the optical matrix element. Because the matrix element consists of a product of a plane wave part that varies slowly over unit cells, and a part that is periodic in unit cells, the product decomposes into a sum and a cell-periodic integral.

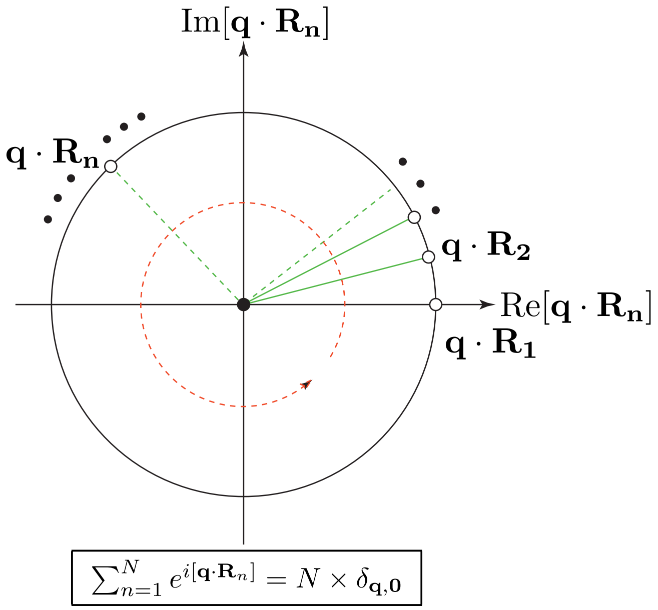

$\mathrm{Im}[\mathbf{q}\cdot\mathbf{R_n}]$
$\mathbf{q}\cdot\mathbf{R_n}$
$\mathbf{q}\cdot\mathbf{R_2}$
$\mathrm{Re}[\mathbf{q}\cdot\mathbf{R_n}]$
$\mathbf{q}\cdot\mathbf{R_1}$
$\sum_{n=1}^{N} e^{i[\mathbf{q}\cdot\mathbf{R}_n]} = N \times \delta_{\mathbf{q},\mathbf{0}}$

Fig. 17.4: The sum of complex exponentials of the form $e^{i\theta_n}$. If the sum is over a large number of phases, the sum $\sum_n e^{i\theta_n}$ is zero, unless $\theta_n = 0$, in which case $\sum_n e^{i\theta_n} = N$. This statement is captured in $\sum_n e^{i\theta_n} = N\delta_{\theta_n,0}$.

for all cases except when $-\mathbf{k_c} + \mathbf{k_v} + \mathbf{k_{op}} = 0$, in which case it is evidently unity. The complex numbers $e^{i\theta_n}$ are all on the unit circle on the complex plane, and if there are a *lot* of them, they distribute uniformly around the origin. Thus, their sum tends to have zero real and imaginary parts; further, they are divided by a large number N. But when $\theta_n = 0$, all the points fall at $e^{i0} = 1 + 0i$, and thus the sum is unity.

The optical matrix element is thus given by the very important result

$$\boxed{W_{ab} = \frac{eA_0}{2m_0}[\delta_{\mathbf{k_c},\mathbf{k_v}+\mathbf{k_{op}}}](\hat{e}\cdot\mathbf{p_{cv}})} \tag{17.42}$$

Note that the Kronecker-delta function ensures momentum conservation because $\hbar\mathbf{k_v} + \hbar\mathbf{k_{op}} = \hbar\mathbf{k_c}$. With this form of the optical matrix element, the net absorption rate from equation 17.28 becomes

$$R(\hbar\omega) = (\frac{eA_0}{2m_0})^2 \frac{2}{V}\sum_{\mathbf{k_c}}\sum_{\mathbf{k_v}}\frac{2\pi}{\hbar}|\hat{e}\cdot\mathbf{p_{cv}}|^2\delta^2_{\mathbf{k_c},\mathbf{k_v}+\mathbf{k_{op}}}\delta[E_c(\mathbf{k_c})-(E_v(\mathbf{k_v})+\hbar\omega)]\times[f_v(\mathbf{k_v})-f_c(\mathbf{k_c})] \tag{17.43}$$

We note that the square of the Kronecker-delta function is the same as the Kronecker-delta $\delta^2_{\mathbf{k_c},\mathbf{k_v}+\mathbf{k_{op}}} = \delta_{\mathbf{k_c},\mathbf{k_v}+\mathbf{k_{op}}}$. We also note at this point that $|\mathbf{k_c}|, |\mathbf{k_v}| >> \mathbf{k_{op}}$. This is because the band-edge states occur around reciprocal lattice vectors $2\pi/a_0$, and the lattice constants $a_0 << \lambda$, the wavelength of light. This is the rationale behind the commonly stated fact: *direct optical transitions are vertical in $E(\mathbf{k})$ diagrams.* Using the Kronecker delta function to reduce the summation over $\mathbf{k}$ states assuming $\mathbf{k_c} = \mathbf{k_v} = \mathbf{k}$, and $\mathbf{k_{op}} \approx 0$, and taking the term $\hat{e}\cdot\mathbf{p_{cv}}$ out of the sum because it does not depend on $\mathbf{k}$, we obtain the net absorption rate per unit volume to be given by the following form, which actually holds also for lower-dimensional structures such as quantum wells, wires, or dots:

$$\boxed{R(\hbar\omega) = \frac{2\pi}{\hbar}(\frac{eA_0}{2m_0})^2|\hat{e}\cdot\mathbf{p_{cv}}|^2\frac{2}{V}\sum_{\mathbf{k}}\delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)] \times [f_v(\mathbf{k}) - f_c(\mathbf{k})]} \quad (17.44)$$

17.6 The Equilibrium Absorption Coefficient $\alpha_0(\hbar\omega)$

We are now ready to evaluate the absorption coefficient. Using the expression for $R(\hbar\omega)$ with the photon flux $N_{ph}(\hbar\omega)$ from Equation 17.23, the expression for the absorption coefficient from Equation 17.13 becomes

$$\alpha(\hbar\omega) = (\underbrace{\frac{\pi e^2}{n_r c\epsilon_0 m_0^2\omega}}_{C_0})|\hat{e}\cdot\mathbf{p_{cv}}|^2\frac{2}{V}\sum_{\mathbf{k}}\delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)] \times [f_v(\mathbf{k}) - f_c(\mathbf{k})]. \quad (17.45)$$

Notice that the absorption coefficient thus formulated becomes *independent* of the intensity of the incident photon radiation $I \propto A_0^2$ because both $N_{ph}(\hbar\omega) \propto A_0^2$ and $R(\hbar\omega) \propto A_0^2$, and the A_0^2 factor thus cancels in the ratio. This is a signature of a linear process - i.e., the linear absorption coefficient of the semiconductor is a property of the semiconductor alone, and does not dependent on the excitation intensity. With the coefficient $C_0 = \frac{\pi e^2}{n_r c\epsilon_0 m_0^2\omega}$ we re-write the absorption coefficient again as the following compact expression which will be used also for lower-dimensional structures such as quantum wells, wires, or dots in chapter 11:

$$\boxed{\alpha(\hbar\omega) = C_0|\hat{e}\cdot\mathbf{p_{cv}}|^2\frac{2}{V}\sum_{\mathbf{k}}\delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)] \times [f_v(\mathbf{k}) - f_c(\mathbf{k})]} \quad (17.46)$$

To evaluate the $\mathbf{k}-$sum, we need to identify the occupation functions $f_v(\mathbf{k})$ and $f_c(\mathbf{k})$. If the semiconductor is in equilibrium, there is one Fermi level E_F, and the occupation is given by the Fermi-Dirac function $f(E) = (1 + \exp[(E - E_F)/k_BT])^{-1}$ at temperature T. When the semiconductor is pushed to non-equilibrium by either optical excitation or electrical injection of excess carriers, the occupation functions are conveniently modeled by retaining the Fermi-Dirac form. But the single Fermi-energy E_F splits to two quasi-Fermi levels: one for electrons in the conduction band F_c, and the other for electrons in the valence band F_v. The occupation functions are then given by

$$f_v(\mathbf{k}) = \frac{1}{1 + \exp(\frac{E_v(\mathbf{k}) - F_v}{kT})} \quad (17.47)$$

$$f_c(\mathbf{k}) = \frac{1}{1 + \exp(\frac{E_c(\mathbf{k}) - F_c}{kT})} \quad (17.48)$$

We will consider non-equilibrium conditions in the next chapter. Under thermal equilibrium, $F_c = F_v = E_F$, and there is only one Fermi level[3]. For an undoped semiconductor, E_F locates close to the middle of the bandgap. Then, as $T \to 0$ K, $f_v(\mathbf{k}) \to 1$ and $f_c(\mathbf{k}) \to 0$. Actually, these conditions hold fine even at room temperature for wide-bandgap semiconductors with little error. Converting the sum to an integral using the usual prescription, we get the *equilibrium* absorption coefficient to be

[3] When photons are incident on the semiconductor, it is by definition not in equilibrium and $F_c \neq F_v$. But we assume that the intensity of the EMag wave is low enough to ensure that $F_c \approx F_v \approx E_F$.

$$\alpha_0(\hbar\omega) = C_0|\hat{e}\cdot\mathbf{p_{cv}}|^2\frac{2}{V} \times \int_{\mathbf{k}} \frac{d^3\mathbf{k}}{\frac{(2\pi)^3}{V}}\delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)] \tag{17.49}$$

Note that the volume term V cancels.

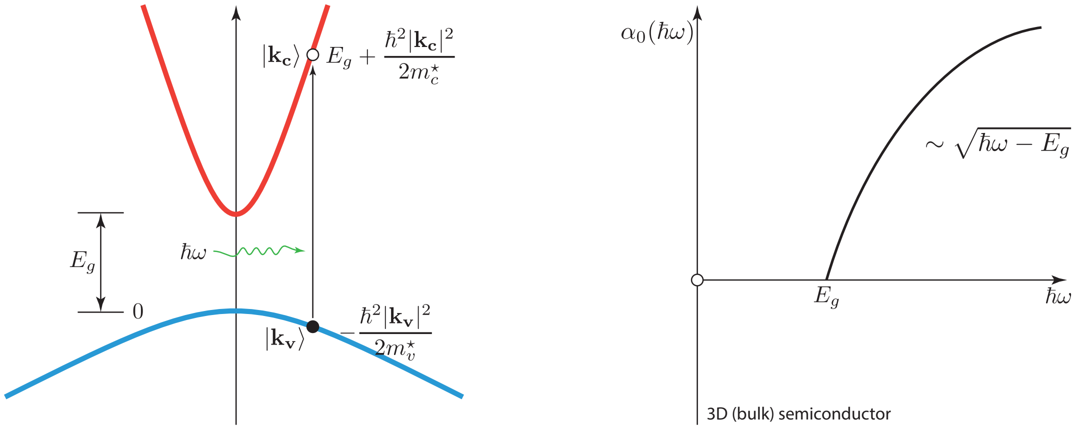


Fig. 17.5: Definition of various energies, and the equilibrium absorption spectrum of bulk (3D) semiconductors $\alpha_0(\hbar\omega)$.

From Figure 17.5, the optical transition can occur only for $\mathbf{k}$ states that satisfy

$$E_c(\mathbf{k}) = E_g + \frac{\hbar^2k^2}{2m_e^\star} \tag{17.50}$$

$$E_v(\mathbf{k}) = -\frac{\hbar^2k^2}{2m_h^\star} \tag{17.51}$$

$$E_c(\mathbf{k}) - E_v(\mathbf{k}) = E_g + \frac{\hbar^2k^2}{2m_r^\star} \tag{17.52}$$

$$\frac{1}{m_r^\star} = \frac{1}{m_e^\star} + \frac{1}{m_h^\star} \tag{17.53}$$

Using spherical coordinates in the 3D $\mathbf{k}-$space, $d^3\mathbf{k} = k^2\sin\theta dk d\theta d\phi$, we convert the variables from wavevector to energy. Assuming $E = \frac{\hbar^2k^2}{2m_r^\star}$, we break up $k^2\sin\theta dk d\theta d\phi$ into three parts: $k^2\cdot dk = (\frac{2m_r^\star}{\hbar^2})E\cdot\frac{1}{2}(\frac{2m_r^\star}{\hbar^2})^{\frac{1}{2}}\frac{dE}{\sqrt{E}} = \frac{1}{2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{E}dE$, the second part being $\sin\theta d\theta$ and the third part $d\phi$. When we integrate over all $k-$space, the angular parts evaluate to $\int_0^\pi \sin\theta d\theta = 2$ and $\int_0^{2\pi} d\phi = 2\pi$.

The absorption coefficient then becomes

$$\alpha_0(\hbar\omega) = C_0|\hat{e}\cdot\mathbf{p_{cv}}|^2\frac{2}{(2\pi)^3}\cdot(2\pi)\cdot(2)\cdot\frac{1}{2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\underbrace{\int_0^\infty dE\sqrt{E}\times\delta[E-(\hbar\omega-E_g)]}_{\sqrt{\hbar\omega-E_g}} \tag{17.54}$$

which reduces to

$$\boxed{\alpha_0(\hbar\omega) = C_0|\hat{e}\cdot\mathbf{p_{cv}}|^2 \underbrace{\frac{2}{(2\pi)^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{\hbar\omega - E_g}}_{\rho_r(\hbar\omega - E_g)}} \tag{17.55}$$

where we have defined the joint optical density of states function for bulk 3D semiconductors as

$$\rho_r(u) = \frac{g_s}{(2\pi)^2}\cdot(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\cdot\sqrt{u} \tag{17.56}$$

Figure 17.5 shows the equilibrium absorption spectrum $\alpha_0(\hbar\omega)$ of a bulk 3D semiconductor. Using typical values of effective masses and material constants, it may be verified that the absorption coefficient for GaN for example are of the order of $\sim 10^5$ cm^{-1}, as indicated in Fig 17.1 at the beginning of this chapter. The absorption coefficient is zero for photon energies below the bandgap of the semiconductor, as is intuitively expected.

Instead of leaving the expression for the absorption coefficient in terms of the unphysical parameter C_0, we use the fundamental Rydberg energy $R_\infty = \frac{e^2}{4\pi\epsilon_0(2a_B)}$, the Bohr radius $a_B = \frac{\hbar}{m_0 c\alpha}$, and the fine structure constant $\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c}$ to write the absorption coefficient as

$$\boxed{\alpha_0(\hbar\omega) = (\frac{4\pi^2\alpha}{n_r})\cdot(R_\infty a_B^2)\cdot(\frac{\frac{2|\hat{e}\cdot\mathbf{p_{cv}}|^2}{m_0}}{\hbar\omega})\cdot\rho_r(\hbar\omega - E_g)} \tag{17.57}$$

where we have split off the dimensionless term $2|\hat{e}\cdot\mathbf{p_{cv}}|^2/m_0\hbar\omega$. Note that as discussed in chapter 9, the rough order of $2|\hat{e}\cdot\mathbf{p_{cv}}|^2/m_0 \approx 20$ eV for most bulk semiconductors. The coefficients decompose to reveal a proportionality to the fine-structure constant. The term $R_\infty a_B^2$ has units eV.cm^2, and the reduced density of states is in 1/eV.cm^3, which leads to the net units cm^{-1}. This general form of the equilibrium absorption coefficient holds even for low-dimensional structures with the suitable DOS $\rho_r(\hbar\omega - E_g^{'})$, where $E_g^{'}$ accounts for ground state quantization shifts in the bandgap. Many interesting effects happen when the semiconductor is pushed out of equilibrium: it is the subject of the next chapter.

Bibliography

[1] C. Wood and D. Jena. *Polarization Effects in Semiconductors: From Ab-Initio Theory to Device Applications*. Springer, New York, 2007.

[2] K. Seeger. *Semiconductor Physics, An Introduction*. Springer Verlag, Berlin, 6th edition, 1999.

[3] C. M. Wolfe, N. Holonyak Jr., and G. E. Stillman. *Physical Properties of Semiconductors*. Prentice Hall, Englewood Cliffs, New Jersey, 1st edition, 1989.

[4] J. H. Davies. *The Physics of Low-Dimensional Semiconductors*. Cambridge University Press, Cambridge, United Kingdom, 1st edition, 1998.

[5] W. B. Joyce and R. W. Dixon. *Appl. Phys. Lett.*, 31:354, 1977.

[6] C Hamaguchi. *Basic Semiconductor Physics*, page 280, 2001.

[7] J. Ziman. *Theory of Solids*, Cambridge University Press, 1972.

[8] D. K. Ferry. *Semiconductor Transport*. Taylor & Francis, London, 1st edition, 2000.

[9] T. Ando, A. B. Fowler, and F. Stern. *Rev. Mod. Phy.*, 54:437, 1982.