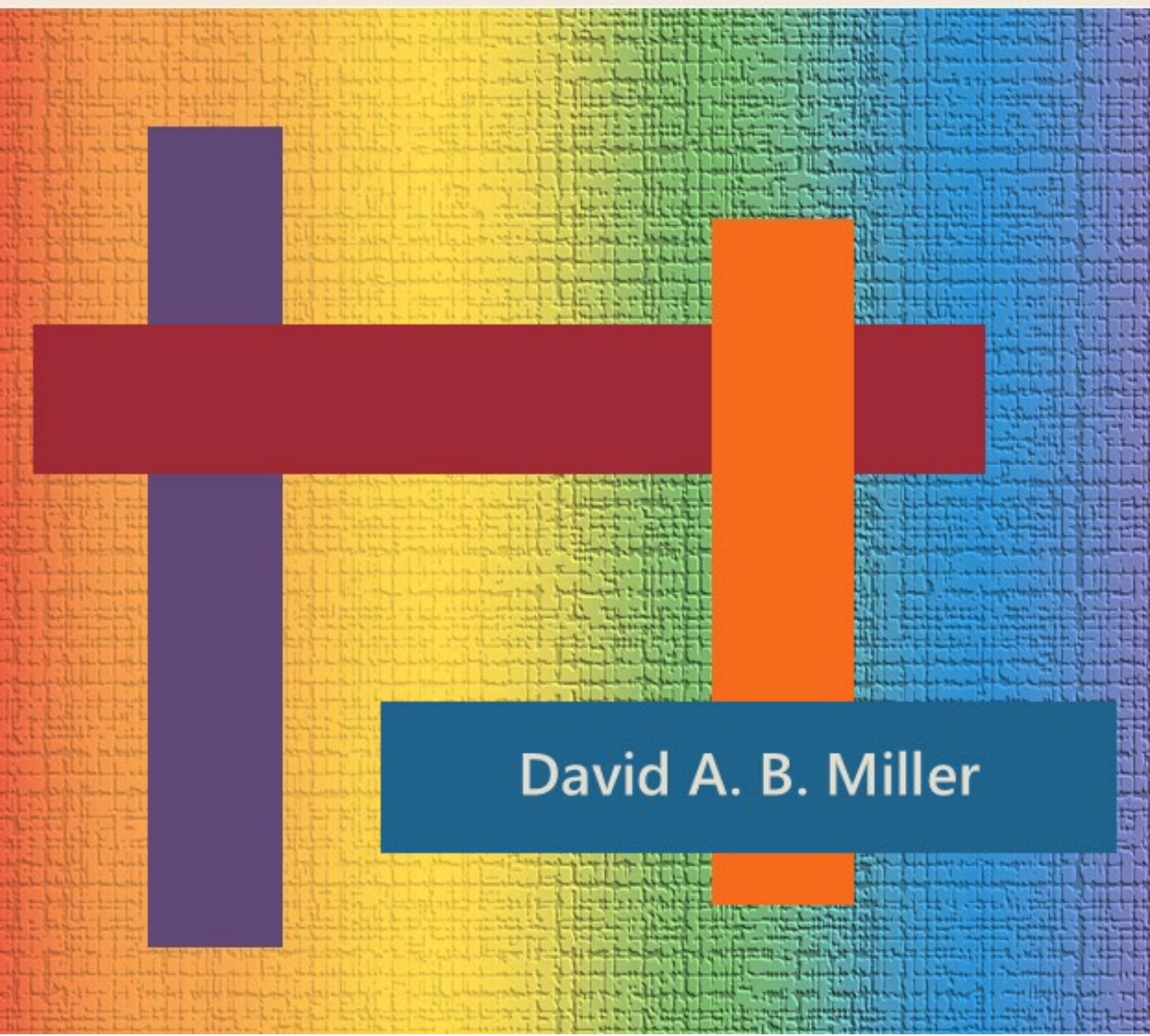


Modern Physics for Engineers and Scientists



David A. B. Miller

Modern Physics for Engineers and Scientists

David A. B. Miller

Stanford University

Copyright © David A. B. Miller 2025.

This work is openly licensed by CC BY-NC-ND 4.0.



See <https://creativecommons.org/licenses/by-nc-nd/4.0/>

ISBN: 979-8-9987410-1-2 (hardback)

ISBN: 979-8-9987410-0-5 (paperback)

ISBN: 979-8-9987410-2-9 (digital)

First published 2025

Imprint: Miller Science

Hardback and paperback versions published through Kindle Direct Publishing (KDP)

In memory of my father, Barclay Miller

Contents

Preface		iv
How to use this book		vii
Prologue		x
Chapter 1	Introduction	1
1.1	Why modern physics?	1
1.2	The background to modern physics	2
1.3	The development of classical physics for engineering	9
1.4	Introduction to the main topics of this book	10
Chapter 2	Oscillations and waves	12
2.1	Introduction	12
2.2	Elementary classical mechanics	13
2.3	Modes	19
2.4	Analysis of systems with modes	21
2.5	The classical wave equation	28
2.6	The Helmholtz equation	32
2.7	Standing waves	36
2.8	A coupled oscillator	40
2.9	Mathematical properties of modes	44
2.10	Conclusions	57
Chapter 3	The quantum view of the world	58
3.1	The beginning of quantum mechanics	58
3.2	The early quantum mechanics of atoms	63
3.3	Schrödinger's wave equation	69
3.4	Probability density	73
3.5	The "particle in a box" model	73
3.6	Waves and measurement	85
3.7	Tunneling	99
3.8	Conclusions	112
Chapter 4	Particles, atoms, and crystals	114
4.1	The hydrogen atom	114
4.2	Center-of-mass behavior	114
4.3	The hydrogen atom Schrödinger equation	116
4.4	Spherical harmonics	120
4.5	Radial solutions and energy levels	127
4.6	Overall form of the hydrogen atom orbitals	132
4.7	Spin, electrons, and Pauli exclusion	134

4.8	Many-electron atoms	135
4.9	Fermions and bosons	140
4.10	Coupled systems	143
4.11	Crystals	148
4.12	Emergence of bands	153
4.13	Band structures in crystals	158
4.14	Band structures in one dimension	166
4.15	Band structures in three dimensions	170
4.16	Electrons in bands	175
4.17	Conclusions	181
Chapter 5	Thermal distributions	182
5.1	Statistics and physics	182
5.2	Tossing coins	183
5.3	Two-state spin systems	190
5.4	Entropy and temperature	196
5.5	Carnot efficiency limit for heat engines	209
5.6	The Boltzmann factor	212
5.7	Chemical potential and the Gibbs factor	215
5.8	Distribution functions	221
5.9	Conclusions	232
Chapter 6	Bands and electronic devices	234
6.1	Introduction	234
6.2	Transport of carriers	235
6.3	Fermi levels and doping	239
6.4	Diodes	243
6.5	Conclusions	253
Chapter 7	Light and quantum mechanics	254
7.1	Introduction	254
7.2	The photoelectric effect	255
7.3	Light and modes	257
7.4	Thermal radiation	259
7.5	Black-body radiation	263
7.6	Absorption and emission of light	266
7.7	Conclusions	271
Chapter 8	Semiconductor optoelectronics	272
8.1	Introduction	272
8.2	Photodetectors and solar cells	273
8.3	Light-emitting diodes and lasers	279
8.4	Conclusions	289
Chapter 9	Epilogue	290

Appendix A	Physical constants and units	292
A.1	Fundamental constants	292
A.2	SI units	293

Appendix B	Greek alphabet	295

Appendix C	Thermal carrier distributions	296
C.1	Introduction	296
C.2	Density of states	296
C.3	Carrier concentrations and Fermi levels	298
C.4	Metals and degenerate statistics	299
C.5	Electrons and holes	301
C.6	Law of mass action and intrinsic carrier concentration	303
C.7	Doping and Fermi levels	306

Appendix D	Quiz solutions	309

Appendix E	Solutions to selected problems	332

Appendix F	Online videos	348
F.1	Modern Physics for Engineers and Scientists videos and slides	348
F.2	Short list of lectures and slides	348
F.3	Full list of lectures and slides	352

Index		375

Preface

The physical technologies we need for key applications of major societal impact, from solar energy to the ways we sense, move, store and process information, are based on ideas and concepts beyond the world of classical physics. A purely classical view gives us no model for the atoms and molecules of chemistry or of material properties such as color or electrical conductivity. We cannot understand the simplest solar cell or even the color and brightness of the sun. To move forward, we need core ideas from quantum mechanics and statistical mechanics. The transistors, lasers, photodetectors, and memory devices that enable all of information technology similarly rely on these same concepts. Without those ideas, we simply cannot understand how this all works, and we cannot make the advances we want.

To be able to create and apply the emerging technology for these and other important societal problems, and to understand the limits imposed by science on the various solutions, the next generation of engineers and scientists needs an understanding and a working knowledge of ideas beyond the classical world. Unless we commit seriously to studying physics itself, it can, however, be difficult for us to acquire that understanding. Typically, we would need many classes just to cover the necessary range of concepts, and we may not have the time (or inclination) for that. Even physics students may find it challenging to get an economical yet substantive introduction to the necessary ideas.

One common approach beyond core classical ideas like mechanics and electromagnetism is to teach a sequence including introductory modern physics, followed by deeper courses in quantum mechanics, thermodynamics and statistical mechanics. For many engineers and for scientists outside of physics, however, there may not be enough time available for such a comprehensive syllabus before we need to start using these ideas.

I have had the opportunity to work in physics and engineering for many years, both inside and outside of academia, and to have taught over this spectrum to a wide range of students in engineering and science at Stanford. This text, and the course “Modern Physics for Engineers” for which it has been developed, are my attempt to solve this problem of introducing the core ideas of modern physics with enough rigor to give a solid basis, but still in an efficient fashion. Indeed, I believe we can do this in a one-quarter or one-semester course by carefully choosing the core ideas and finding the most economical yet coherent path through them. I believe too that the interested student can use this book and its many resources to teach themselves this core modern physics.

Obviously, we cannot go as deeply into many topics as such a longer suite of courses could allow. By teaching an introduction in our one coordinated approach, however, we have one major advantage: we can teach the topics in the sequence that makes the most intellectual sense, saving ourselves considerable time and effort. That sequence is

different from the historical order in which these areas emerged and in which they are sometimes taught.

Here we start with the ideas of modes from classical oscillators and waves. These ideas, which are sometimes not covered sufficiently in introductory classical physics, are useful in themselves; more importantly here, however, they set the intellectual ground for the concepts of quantum mechanics, with its extensive use of the ideas of eigenstates (which are, mathematically, essentially the same concept as modes).

Building on these mode and eigenstate concepts, we can then introduce quantum mechanics. After presenting the key ideas of the quantum mechanical view of the world, we can proceed to atoms and crystals. The understanding of quantum mechanics in crystals is directly useful for a wide range of applications, and gives a first example of a quantum mechanical treatment of the properties of materials.

The emergence of states in quantum mechanics sets us up to discuss the ideas of the statistical occupation of these states, which in turn allows us to introduce the ideas of entropy and the various statistical distributions for occupations of the states. Then we can introduce some key ideas of thermodynamics, though we can largely avoid having to introduce the rest of the bulk of that substantial subject at this point.

With this background in place, we can present some core applications in the physics of electronic devices. After this, we give a first treatment of the quantum mechanics of light and its absorption and emission. This enables a short introduction to optical and optoelectronic devices, including those used in optical communications and solar cells. The goals in briefly introducing device applications are to give some applied context to the basic scientific ideas, showing how those ideas impact real engineering, and to leave students well prepared if they choose to take further device courses. These device applications also give good examples to exercise the key ideas.

The ideas we introduce here really do allow us to understand much that a classical view simply cannot explain, and they give us a sound basis of concepts for engineering modern technology in a wide range of applications. Though we cannot teach everything in this text, I believe we can teach enough to give the reader a strong basic understanding of how the modern world actually works. Indeed, we can see that there is a relatively small set of physical processes that underlie the modern physics for a broad range of applications. These ideas are also themselves intellectually fascinating and stimulating, and they change the way that we think about the world.

Some background in mathematics and physics is necessary for working through the topics in this book, though I try to minimize this background as much as I can. I presume the reader knows calculus through introductory differential equations and has basic understanding of matrices and vectors and some associated linear algebra. I introduce any other mathematics we might need beyond this, and I try not to make this book any more mathematical than it needs to be. I presume also that the student has had basic physics classes, but I try to review most of the physics we need (and I do not presume the student is necessarily a physics major!).

I am pleased here to acknowledge the many contributions of Austin Brownlow, Christine Donnelly, Ching-Ying Lu, Matthew Morea, and Marina Radulaski who worked with me in teaching this class as it was being developed, and correcting many

of my errors, and I'm grateful to the insights of my professorial colleagues Dan Congreve and Tony Heinz in also teaching this class. Finally, I am very happy to acknowledge my wife Pat, both for proof-reading this text (though the responsibility of all errors remains mine!) and for tolerating me during the writing of this text (and at other times too!).

David Miller

Stanford, April 2025

How to use this book

Who this book is for

The intent of this book is to introduce these key ideas of modern physics to anyone with a moderate college science or engineering background. This book has been designed both as a classroom text and as a good way for any such interested student to learn this material themselves.

The material is at a level accessible for students studying science or engineering from second year college or beyond, and could be suitable also for a well-prepared first-year student. The title “Modern Physics for Engineers and Scientists” is chosen in part to reassure engineers and scientists outside of physics who might otherwise be (or already have been) discouraged from such a topic. The material is, however, suitable for scientists and engineers in many fields, including physics itself.

The main adjustments compared to some standard physics sequences are, first, that the material deliberately minimizes the required physics to start the course and, second, that the material is introduced in the order that makes most intellectual sense, rather than in historical sequence. As a course, it is intended to be particularly economical and efficient in introducing its core topics of quantum mechanics and statistical mechanics, hence allowing it to be included in a syllabus of courses that will have many other demands on student learning time. No matter what your scientific or engineering background or interest, this book offers a first introduction to these topics that is hopefully rigorous, accessible, and, hopefully without undue optimism, even enjoyable!

Teaching and learning approaches

The material is sufficient for a full “quarter” (e.g., 10 week at ~ 3 lectures per week) course, or, at a more relaxed pace, a “semester” (e.g., 15 week at ~ 2 lectures per week) course. As such, it can form the basis for a conventional lecture course, for which more than sufficient problems are included for weekly problem sets throughout the course.

Sequence of topics

The material is presented in a coherent sequence.

- 1) A short background history of physical science up to about 1860 (Chapter 1)
- 2) Background physics and mathematics (Chapter 2)
- 3) An introduction to quantum mechanics (Chapter 3) and its use to understand atoms and crystalline materials (Chapter 4)

- 4) An introduction to statistical mechanics, how it underpins key thermodynamic ideas, and how it gives thermal statistical distributions of particles like electrons and photons (Chapter 5).
- 5) Use of the quantum mechanics and statistical mechanics to understand the physics of electronic devices (Chapter 6). Appendix C also extends this discussion for those particularly interested in this topic.
- 6) The basic quantum mechanics of light, including thermal radiation and absorption and emission (Chapter 7).
- 7) The use of these ideas for understanding semiconductor optoelectronic devices (Chapter 8).

Using parts of this sequence

Parts of this sequence can be used on their own, and some topics can be omitted. Chapters 3 and 4 (supported by Chapter 2) give a solid first introduction to quantum mechanics. Chapter 5 can be used on its own as a first introduction to statistical mechanics. If the material in Chapters 2-5 is covered, Chapter 6 (and optionally Appendix C) can be used for understanding electronic device physics. Again following Chapters 2-5, Chapter 7 can introduce the quantum and statistical mechanics of optics and optionally Chapter 8 can introduce optoelectronic device physics.

Special features of this book

The material has three additional features – quiz questions, selected problem solutions, and online videos – that can help in class-room teaching or with other or extended ways of learning.

For students studying on their own, hopefully these three features make it easier to work through the course, in whatever style the student prefers. These features can also be useful in “flipped classroom” teaching approaches that emphasize the use of class time for problem solving and discussion. The students could, for example, watch the online videos before the class session. Class time could also make use of the quiz questions to stimulate discussion (possibly with “closed books”), or go one to solve some of the problems.

Quiz questions

In the book itself, there are extensive “quiz” questions at the end of most sections. These quiz questions require little or no calculation or algebra to answer, so should take little time. Rather, they allow the student to test and exercise their conceptual understanding of the material just presented. Importantly, *all* the answers to these questions are available in Appendix D.

Solutions to selected problems

Full solutions to selected problems are provided in Appendix E. This will be especially useful for students studying the material on their own, and can be an additional resource in class-room teaching, providing worked examples.

Online lectures

Perhaps most unusually, there is a complete set of 25 online lectures, all openly available, that present all the material in the main text (Chapters 1 through 9), in order. These lecture videos are presented in multiple shorter sections (to help with our finite concentration spans!).

An icon, such as  shows the point in the text corresponding to the start of each numbered lecture section. These icons are live in the digital version of the text, and clicking on them will bring up the lecture video.

A complete list with links to videos and slide copies and references to the relevant text sections is provided in Appendix F. This is also available as a pdf document with live links through [this link](https://purl.stanford.edu/vg346xh7067) <https://purl.stanford.edu/vg346xh7067> or the QR code on the right.



Openly available material



In addition to conventional paper book versions, this entire book is openly and freely available in digital form. See the link <https://purl.stanford.edu/hr256mp8317> or the QR code on the left.

In either paper or digital form, the material is copyrighted but is licensed under the CC BY-NC-ND 4.0 license. (See <https://creativecommons.org/licenses/by-nc-nd/4.0/> for details.) This license enables reusers to copy and distribute the material in any medium or format in unadapted form only, for noncommercial purposes only, and only so long as attribution is given to the creator.

Prologue

There are two pieces of good news about understanding the everyday physical world around us. First, we largely know how it works – we can write down some principles that let us describe essentially all the physical behaviors we see, from the inside of atoms to the outside of the solar system. Second, there are not that many principles we need to learn to do this. That might be a surprise, because there is certainly enough science to fill many libraries. But a few concepts well understood can go a very long way.

There are real bounds to this conceit, of course. Complex systems still confound us; to take two examples, it is hard for us to predict the weather because the calculations are too hard, and our brains still do not understand how they themselves work. Nonetheless, we still likely presume that neither the behavior of the weather nor the processes in our neurons are violating our physical principles. We should, of course, be honest that there are aspects, especially of quantum mechanics, that puzzle us even though we understand in a practical sense how to use them. And, we realize there is much we do not know when we head off to cosmological scales or back to the big bang or down to the depths of elementary particles. Despite these caveats, our understanding of some key physical ideas allows us to engineer much of the material world and the energy, transportation, communication, and information that help us run it.

The challenge for understanding our physical world is that it does not actually work the way we might think. The “classical” view of the simple mechanics of objects that we learn in elementary physics classes, while useful in many ways, leads us to dead ends if we ask questions like why materials have chemical properties, or why the sun is the color it is, or how to design a better transistor or solar cell. Unless you have had the opportunity or even the luxury to study quite a lot of physics, you are unlikely to have been able to “peer behind the curtain” to reveal what is going on. You will not have encountered many of the principles we need, let alone see how they fit together into a coherent picture.

At this point, you may think hope is lost; you do not have the time (or perhaps the inclination) for that deep and comprehensive study of half a dozen areas of intimidating physics. And, it is true, as it is for most worthwhile endeavors, that some effort and persistence is going to be required here. But, there are three other pieces of good news.

First, it turns out that much of the physics we need was discovered in the wrong order. That itself is not good news, of course. But, by putting it together in the right order, we can save a lot of time while also developing a coherent story.

Second, though we cannot and will not fully develop the various areas of physics we introduce, with just a moderate amount of time and effort we can put together a coherent and compact core. We can progressively grow the “trunk” of a tree of concepts here, even though we will only be able just to start some of the branches that it can support.

Third, because the principles we learn are so economical yet universal, and often quite unexpected, encountering them can truly be an exciting intellectual journey where we are drawn forward by successive and quite remarkable revelations.

So, fortified by all this good news, we can set out on our adventure to find out how the world actually works. I honestly believe you will find the journey fun and the destination worth the effort.

Chapter 1

Introduction

1.1 Why modern physics?

While we might accept that modern physics is important to modern science, it might be easy to imagine that engineering does not need it. Is modern physics not some esoteric subject concerned only with elementary particles and strange aspects of quantum mechanics remote from everyday existence? To be fair, we do need to be clear what we mean by “modern”. There certainly is a “classical” physics of Newton’s Laws explaining mechanics and moving objects and Maxwell’s equations describing electromagnetism, radio waves and many aspects of light, and we might add the thermodynamics of steam engines and the like into that classical world. There is no doubt we have constructed much useful engineering out of such classical ideas. We could draw the line between classical and modern at roughly that point, which would put us somewhere about three-quarters of the way through the 19th century. Is there really anything important for engineering that we would miss if we stopped there?

Of course, the answer is quite certainly yes – we would be overlooking much of what makes our modern world possible. Stopping at that point would mean we would understand nothing about the physical origin of the properties of materials or of the actual processes of chemistry. We would have no model of the physical basis of color and how light interacts with anything. We would have no idea how light bulbs actually work, or why the sun gives off the amount of light it does, or even why the sun is the color it is. Why some materials conduct electricity and others insulate would be a complete mystery.

Without the modern physics beyond that classical boundary at about 1870, we would have none of the devices of the information age – no transistors, no lasers, no digital cameras, no optical fiber communications, no televisions, no mobile phones, no computers, no display screens, no hard drives or any other digital memory; we would not even have vacuum tubes. There would be no solar cells and no LED (light-emitting diode) lighting to tackle our energy problems. Without the modern physics past our boundary at ~ 1870, we would have little of what makes modern technology “modern”.

1.2 The background to modern physics

To understand just what we mean here by modern physics, we can start by looking at physical science and how it got to where it was towards the end of the 19th century. This will also help review the background science we need so we can move forward.

Matter

It may seem odd to us now, but there was a time before our ancestors seriously considered the idea that everything was made of something. In that simpler view, a rock is just a rock, a tree is just a tree, water is just water. Perhaps in those times we just had mythological explanations of how things came to be. Now our elementary science education teaches us that all the matter we see in our everyday lives is made up from atoms, which are themselves made from electrons, protons and neutrons. But, that is a very modern notion, one that did not fully coalesce until several decades into the 20th century.

We can see the progression of thought towards a scientific view in the thinking of the ancient Greek philosophers. The idea that matter is made of something dates back at least to Thales of Miletus (c. 620 BCE – c. 546 BCE). Empedocles (c. 490 – c. 430 BCE) proposed the idea of earth, air, fire and water as being the four elemental substances that made up matter, an approach later also favored by Aristotle (384 – 322 BCE). Though Democritus (c. 460 – c. 370 BCE) and/or his teacher Leucippus proposed that matter was composed of indivisible “atoms”¹, the “four elements” idea would remain very influential for nearly 2000 years. (Today, we would recognize these ideas as being closer to a statement of the different phases of matter – solid (earth), liquid (water), gas (air), and fire (plasma) rather than matter’s constituents.)

This “four elements” approach formed the basis of alchemy, the precursor of modern chemistry. There were some notable additions such as the elements sulphur and mercury (or at least the philosophical essences of dry exhalations (sulphur) and moist exhalations (mercury)); these were discussed, together with other chemical ideas, in works attributed to Jabir² in about the 700’s CE. The “four elements” idea was also not universally accepted during this time, being disputed by some, such as Avicenna³ (c. 980 – 1037).

The period from about 1500 – 1660 was a transitional one from alchemical ideas towards a modern atomist approach, and to the use of what we would call the scientific method. The ideas of the scientific method, involving a cycle of hypotheses, experiments to test them, modified hypotheses, and so on, has roots extending back to the deductive methods of Parmenides (c. 515 – c. 460 BCE), Leucippus and Democritus, and through scientific experiments such as those by Alhazen (Ibn al-

¹ “Atom” is from the Greek, meaning “cannot be cut”.

² Abu Mūsā Jābir ibn Hayyān, fl. c. 721 – c. 815, also known as Geber

³ Ibn-Sīnā, full name Abū ‘Alī al-Ḥusayn ibn ‘Abd Allāh ibn Al-Hasan ibn Ali ibn Sīnā

Haytham)⁴ (c. 965 – c. 1040). What we would recognize as modern ideas of scientific method were discussed by Francis Bacon, René Descartes and Galileo in the early 1600's.

By the time we get to Robert Boyle (1627–1691), we find a rigorous experimental approach towards chemistry. In 1661, in “The Sceptical Chymist” he argued towards modern ideas that we would think of as atoms, molecules and chemical reactions. He also argued for chemistry to be regarded as what we would call a science, distancing it from some of the philosophical aspects of alchemy.

The subsequent period through the 1700's saw progressive advances in chemistry, with identification of various elements, the emergence of some modern chemical nomenclature, quantitative understanding of reactions, and the idea of conservation of mass in chemical reactions. Such ideas are found, for example, in Antoine Lavoisier's classic “*Traité Élémentaire de Chimie*”, a key chemistry text published in 1789.

With the observation that compounds are made of definite proportions of their constituents, first by Joseph Louis Proust in 1799, and then extended by John Dalton to show integer ratios⁵ in different compounds, the ground was prepared for a clear atomic theory as proposed by Dalton in the first decade of the 1800's. Though the atomic theory was not fully accepted for about a further 100 years, it proved a very successful basis for chemistry. In 1869 Dmitri Mendeleev published his first periodic table of the elements, of which 66 were known by that time, organized by their relative atomic masses and by their similarity of properties.

But we had no idea of what atoms were, what were the actual differences between atoms, and why they had their specific chemical properties. The full understanding of this atomic view would take at least another 50 years or so, until the creation of modern quantum mechanics.

Laws of motion

Early ideas from Aristotle's time held that, if you did not keep pushing something, it would stop moving. The apparent continuous movement of the “heavens” – the stars, the planets, the sun and the moon – required some different approach, such as saying they lay on perfect celestial spheres centered round the earth. Ptolemy (c. 90 – c. 168 CE) put together a model on such a basis, and was able to calculate astronomical tables. At that time, and even well into the early modern era, because astrology was regarded with some seriousness for predicting events, such tables were considered important. Astrology and astronomy were not really separate subjects, and arguably astrology essentially funded accurate astronomical measurements. Of course, astronomy based on an earth-centered view becomes quite complicated, especially for predicting the paths of planets.

Nicolaus Copernicus (1473 – 1543) realized that the whole subject became much simpler if we viewed the earth and the planets as orbiting the sun, and he published this

⁴ Abū 'Alī al-Ḥasan ibn al-Ḥasan ibn al-Haytham

⁵ We would now refer to such integer ratios as “stoichiometric” or “stoichiometric numbers”.

rather heretical “heliocentric” view just before his death. Copernicus’s realization is often cited as the start of modern science, based on observation and hypothesis that should agree with one another. Other astronomers, notably Tycho Brahe (1546–1601), followed with further accurate measurements. Using Brahe’s data, Johannes Kepler (1571–1630) deduced rules (Kepler’s Laws of Planetary Motion, published in 1609 and 1619) for the elliptical orbits of planets (an improvement on Copernicus’s circular orbits) round the sun, which is then positioned at one of the foci of the ellipses. The dynamics of earth-bound motion meanwhile was receiving more attention. Galileo made his famous observations that a pendulum swung with essentially the same period regardless of the amplitude of its oscillation, argued that velocity of falling bodies was not proportional to their weight, and advocated a law of inertia whereby bodies kept on moving even if they were not explicitly pushed. (This law, with some refinements, essentially later turns into Newton’s Second Law of Motion.)

Then Isaac Newton (1642–1726/7) made the proposal in 1687 of his three Laws of Motion and his gravitational theory, which together were able to explain Kepler’s Laws. (Along the way, he also created calculus to construct his theory.) This was a remarkable success, of course, that arguably set up modern science as a search for rigorous, predictive laws. It can also be argued that it represents a break point from the previous approaches to explanations of the natural world. Prior to Newton’s proposal, it was common to argue that there must be some harmonious plan behind nature, and the behavior we see should follow from that plan⁶. It would be common at that time to regard physical theories with suspicion if they did not correspond to such philosophical plans. Possibly we should even be able to deduce the plan on the basis of pure logic, without making measurements. Newton’s attitude was apparently that it was not necessary to have such a plan for the theory to be correct.

Newton’s mechanics remained the foundation for the dynamics of motion for over 200 years. It is still the method we turn to first for most such calculations in engineering or science today, at least for objects large enough to see with the naked eye or through microscopes and traveling at velocities much less than the speed of light. More mathematically sophisticated versions were developed in the 1700’s and 1800’s, notably the approach by Joseph-Louis Lagrange (1736–1813) in 1788, and extended by William Hamilton (1805–1865) in 1833; with such approaches, we came to view dynamics more in terms of conservation laws, like conservation of energy and of momentum, but the basis of our dynamics remained Newton’s Laws at least till the 20th century.

The theories of relativity introduced by Albert Einstein (the special theory, introduced in 1905 and the general theory, introduced in 1916, which is also the modern theory of gravitation) formally superseded Newton’s models of classical dynamics, though

⁶ Even Kepler, for example, initially considered that the fact there are only five “platonic solids” (tetrahedron, cube, octahedron, icosahedron, and dodecahedron) to be compatible with the fact that there were (apparently at his time) only six planets, with those solids occupying the spaces between the planets, and he continued to look for “harmonious relations” in planetary behaviors throughout his career.

Newton's approaches remain practically useful and valid in most everyday engineering situations for macroscopic objects.

Light, electromagnetism and Maxwell's equations

Light

Lenses and mirrors were well known in ancient times. Some early theories of light, such as advocated by Empedocles, argued that light rays emanated from the eyes towards objects⁷. Sometime around 300 BCE, Euclid constructed a geometrical theory of optics based on light rays traveling in straight lines, which allowed a theory of perspective. Later, Alhazen analyzed the Greek optical theories, but proposed that light travels in rays to the eye from different points on an object, and models by Ibn Sahl⁸ (c. 940–1000) successfully analyzed both reflection and refraction⁹. Wearable eyeglasses were in use by the 1300's.

The modern law of refraction, known as Snell's law, was found by Willebrord Snellius (1580–1626) in 1621. Clear observations of diffraction of light¹⁰ were made in the 1600's by Francesco Grimaldi (1618–1663). Newton's experiments with prisms around 1670–72, splitting up white light and putting it back together, showed that white light is made up out of the different “spectral” colors of light, and led to our modern understanding that objects mostly¹¹ do not generate new colors of light that are not there in the light shining on them. Specifically, in the case of the prism, the prism itself is not generating the new colors, but is instead splitting up the component colors that already exist in the white light.

Newton advocated a “corpuscular” theory of light, with particles of light (corpuscles) traveling in straight lines. Christiaan Huygens in his “*Traité de la lumière*” (Treatise on light) (1690) proposed that light is a wave phenomenon, describing what is known as

⁷ Such an approach does, however, have difficulty explaining why we cannot see in the dark!

⁸ Abu Sa'd al-'Ala' ibn Sahl

⁹ Refraction is the process of the “bending” or change of direction of light rays when they enter materials of different refractive index, and refractive index is the factor by which light travels more slowly in a material compared to its speed in a vacuum.

¹⁰ Diffraction is the “breaking up” of light that we now understand in terms of interference effects of light waves as they encounter edges, small apertures, periodic structures like gratings, or structures that vary on a scale comparable to the wavelength of the light, such as holograms.

¹¹ Objects can luminesce, giving off light, for example when they are hot, and objects can also fluoresce, in which shining light on them at one wavelength (usually a shorter one) leads to light being emitted at another wavelength (usually a longer one). Though fluorescent effects are quite common in certain modern man-made materials, such as some dyes, strong fluorescent effects are relatively rare in nature. Modern non-linear optical techniques can also generate new wavelengths of light from materials, but those typically require the high intensities of lasers to demonstrate them.

Huygens' Principle¹². At that time, there was no direct evidence of this wave nature, and Newton's corpuscular theory appears to have dominated for some time. However, the two-slit experiment on diffraction by Thomas Young (1803) provided conclusive proof for the wave nature of light¹³. Augustin-Jean Fresnel provided a more extensive wave theory of light and diffraction (1821), and with François Arago he determined that light was a "transverse" wave, with polarizations (directions of the electric field vector that are perpendicular to the direction of propagation). Hippolyte Fizeau made the first time-of-flight measurements of the velocity of light in 1849, obtaining an answer within 5% of the modern accepted value. At this point in the mid 1800's, though, there was still no clear connection between light and electromagnetism.

Electromagnetism

The existence of electrostatics and magnetism has been known for a long time; both may have been known to Thales in the 6th century BCE, for example. The ancient Greeks were aware that rubbing amber with fur could lead to attraction between the two, which we now understand as electrostatic attraction. Lodestone (naturally occurring magnetite) was also known to attract iron objects, as was apparently known in the 4th century BCE writings of Wang Xu. The Chinese scientist Shen Kuo (1031 – 1095) later wrote about the magnetic needle compass; writings around 1111 – 1117 by Zhu Yu document its use for navigation. In 1600, William Gilbert concluded that the Earth was magnetic, and that this was the source of the action of magnetic compasses.

The invention of the Leyden jar capacitor (apparently independently by both Pieter van Musschenbroek of Leyden and by Ewald von Kleist in 1745) allowed electrical charge, for example from frictional generation, to be accumulated and stored. Benjamin Franklin flew a kite into a thunderstorm in 1752, and captured the resulting charge from the lightning in a Leyden jar, establishing the link between lightning and electricity. The inverse square law of electrostatics was introduced by Charles Coulomb in 1785. A major discovery and substantial practical step was the creation, by Alessandro Volta in 1799, of the voltaic cell and the battery, using copper or silver discs separated from zinc discs in brine (salt water). This allowed much more convenient experiments using electrical currents and voltages.

Up to this point, electrostatics and magnetism were largely separate phenomena scientifically, but Hans Christian Ørsted in 1820 noticed that a compass needle could be deflected by passing a current through a wire, noting too that the direction of deflection depended on the direction of the current. These observations essentially started the field of electromagnetism. Following this observation, in the same year of 1820, André-Marie Ampère showed that passing a current through a coil of wire caused

¹² Huygens' principle states that we can find the location of the next wave front by considering a set of sources of spherical waves on the current wave front. With some amendments, mostly to get rid of the backwards waves that would result from a literal use of this approach, this becomes quite a good model for wave propagation, including diffraction effects. For a modern version, see D. A. B. Miller, "Huygens's wave propagation principle corrected" *Optics Letters* **16**, 1370-1372, (1991) <https://doi.org/10.1364/OL.16.001370>.

¹³ This approach also allows us to measure the wavelength without having any measuring device on a scale as small as the wavelength.

it to behave like a magnet, and he developed the theory of the magnetic attraction and repulsion of current-carrying wires.

Michael Faraday observed electromagnetic induction in 1831¹⁴. Turning on and off a current in one coil of wire can induce a current in another coil, and moving a magnet through a coil of wire can also similarly induce a current. Hence this Law of Induction shows that changing magnetic fields can create electric fields.

Maxwell's equations

In 1865, James Clerk Maxwell took one more theoretical step, proposing that just as changing magnetic fields could produce electric fields, so also changing electric fields would produce magnetic ones. (The effect of these changing electric fields can be characterized as an effective current, the “displacement current”, which can then be regarded as generating the magnetic field.) With this final step, he was able to synthesize essentially all of electromagnetism in his equations. With changing magnetic fields giving electric fields and changing electric fields giving magnetic ones, these equations predict wave motion. The velocity of that wave motion as calculated from Maxwell's equations essentially agreed with the measured velocity of light, leading to the proposal that light was electromagnetic radiation. Maxwell's equations stand to this day as one of the most remarkable achievements of theoretical physics.

Thermodynamics

Fire has been used since prehistoric times for cooking, heating, and possibly also for clearing land for agriculture. Hero of Alexandria (c. 10 – c. 70 CE) described a simple steam engine, the aeolipile, based on jets of steam from water inside a heated sphere, making this simple steam turbine possibly the first “heat engine”, and he and others may have understood the expansion of air when it is heated. By the early 1600's several scientists¹⁵ were apparently exploiting the idea of a tube closed at one end and with its other end in water as a thermometer. The tube is partially filled with air, and the level of water in the tube varies as the air expands and contracts with changing temperature, an idea that may have been known to Philo of Byzantium (ca. 280 – ca. 220 BCE). Simple steam turbine engines based on steam pushing vanes of some kind were described in 1551 by Taqi al-Din¹⁶, in 1629 by Giovanni Branca, and by John Wilkins in 1648.

The growth of mining required a method to drain the mines of water. Thomas Savery invented the first commercially-used steam engine in 1698, a steam-powered pump. Thomas Newcomen's engine of 1712 employed a piston in a cylinder. The cylinder was filled with steam (not at high pressure), then sealed off and the steam condensed by cooling it with a spray of cold water, so atmospheric pressure would push the piston (from the outside) to give the engine action. James Watt in 1774 invented the separate

¹⁴ Some aspects of this may have been anticipated also by Francesco Zantedeschi in 1829 and 1830.

¹⁵ Possibly Cornelis Drebbel, Robert Fludd, Galileo and Santorio Santorio.

¹⁶ Taqi ad-Din Muhammad ibn Ma'ruf ash-Shami al-Asadi (1526–1585)

condenser steam engine, which greatly improved the efficiency. The importance of efficient steam engines for powering the industrial revolution gave a strong practical need for a deeper understanding of thermodynamics.

A reliable temperature scale and thermometer based on the expansion of mercury was developed by Daniel Fahrenheit in 1724. Joseph Black introduced the idea of latent heat¹⁷ – the heat required to change the phase (e.g., from solid to liquid or from liquid to gas) – in 1762¹⁸. The ice-calorimeter used by Lavoisier in 1782-83 to measure the heat generated from various chemical reactions was based on Black's idea of latent heat. With meaningful thermometers and calorimeters, thermodynamics could start to become a quantitative science.

Sadi Carnot¹⁹ in 1824 considered the question of the efficiency of heat engines like the steam engine, and introduced the idea of an ideal engine and its (“Carnot”) cycle of operation, an engine presumed to be otherwise as good as possible, i.e., having no friction and no conduction of heat other than through the working “fluid” of the device. He proposed that only the difference in the temperature of the “hot” and “cold” reservoirs mattered in such an ideal engine, not the specific “fluid”. At that time, the nature of heat was still not clear, being typically viewed in terms of “caloric” – a supposed fluid that flowed from hot objects to cold objects; this “caloric” was not yet closely identified with energy as we understand it.

A major challenge to the caloric theory was the observation by Benjamin Thompson (Count Rumford) in 1797 that heat could be generated by friction; specifically, by immersing a cannon barrel in water and then attempting to bore it out with a blunt tool, he could boil the water. Hence, mechanical energy could be converted to heat. Using a falling weight to drive a paddle immersed in water, in 1845 James Joule deduced a consistent specific heat for water (the amount of energy required to heat a known mass of water by a known temperature)²⁰. This led to the proposal that in fact heat was energy, and the energy was conserved overall in thermal phenomena (which becomes the First Law of Thermodynamics).

Two other key ideas that emerge from these advances in thermodynamics are entropy and the Second Law of Thermodynamics. The idea of entropy is usually credited to Rudolf Clausius from his work in the 1850's and 1860's. Both Clausius (in 1854) and William Thomson (Lord Kelvin) (in 1851) give statements of the Second Law. The

¹⁷ Since this process occurs at a constant temperature, this heat is hidden – latent – from a thermometer.

¹⁸ Joseph Black, who also discovered carbon dioxide, knew James Watt well, and invested in his company. Both men were very well acquainted with James Hutton, arguably the first modern geologist, and Adam Smith, the progenitor of modern economics. Black, Hutton and Smith were founding members (Fellows) of the Royal Society of Edinburgh in 1783, and Hutton successfully proposed Watt for Fellowship in 1784.

¹⁹ Nicolas Léonard Sadi Carnot (1796-1832), often known just as Sadi Carnot.

²⁰ Similar ideas were being proposed at the same time by Julius von Mayer and Ludwig Colding, though Joule's experiments may have been more precise.

entropy change ΔS resulting from a flow of heat (energy) ΔQ into a system at temperature T in thermodynamics can at least be defined through the simple equation

$$\Delta S = \frac{\Delta Q}{T} \quad (1.1)$$

Here T is expressed relative to absolute zero temperature (a notion introduced by Kelvin in 1848). There are many statements of the Second Law of Thermodynamics, but the key notion, stated somewhat informally, is that, in some isolated system, total entropy cannot decrease. This principle puts a limit on the efficiency of heat engines, for example.

The idea of entropy is one of the subtler ones in physics. Just what entropy means and where the core ideas come from becomes much clearer once we make the change to thinking about kinetic theory – the idea that heat is really the energy stored in the random motions of atoms or other entities. The branch of physics in which we can talk clearly about these ideas is statistical mechanics, which we can regard as being started by James Clerk Maxwell in 1871 and a particularly important paper by Ludwig Boltzmann (1844 – 1906) in 1875, with major contributions by Josiah Gibbs (1839 – 1903).

1.3 The development of classical physics for engineering

Exploiting classical physics

So far, our discussion has mostly taken us up to about 1870, which is a convenient starting point for the major topics of this book. Of course, we can see that there is no real break point here, about three-quarters of the way through the 19th century. The progress of science and engineering was quite continuous and moving at a remarkable pace. Our choice of this point to start “modern” physics is arbitrary, and is dictated more by the topics we want to introduce than by any definite transition. Traditionally, though, at least the physics before this point is regarded as “classical”.

Certainly, in 1870, quantum mechanics as we know it is still some way off. We might reasonably argue that chemistry, with its periodic table of the elements, was a kind of proto-quantum-mechanical science, of course. Anyway, we will mostly have to leave chemistry at this point to focus on the physics that will be the focus of this book.

This classical physics and the engineering based on it continued to develop. In electromagnetism alone, we had the development of electrical power and light, continued extensions of electrical telegraph cables round the world, and the invention of the telephone and of radio transmission. The descendants of these classical technologies form a large part of electrical engineering to this day.

The thermodynamics of heat and entropy, and related concepts like free energy, continued to develop, especially through the work of Josiah Gibbs, laying the basis for the modern chemical thermodynamics that underlies much of chemical engineering, and

underpinning the continued growth in efficiencies and types of heat engines (including internal combustion engines).

There is no doubt that this classical physics was both very powerful and useful, and it continues to be exploited very effectively in many areas of technology today. But, our focus here will now have to move to very different concepts of modern physics, ones that enabled completely new technologies that transformed our world yet again.

1.4 Introduction to the main topics of this book

By about 1870, then, we had much of classical physics in place, but many areas remained completely unexplained. The ideas of chemical elements were known and useful, but we had no underlying theory of them, and no theory of the properties of materials. We understood light, electromagnetism, and waves, but we had no model of how light interacted with anything. We had working concepts for thermodynamics, such as energy, entropy, heat engines, but no clear idea of their physical basis. Our main goal in this book is to introduce those key ideas that come after the “classical physics” of 1870 or so, ideas that underlie the devices of the information age and topics like solar energy. Though there is a very large amount of modern physics after 1870, fortunately there are really just two key areas that underlie our engineering interest here. Those areas are

quantum mechanics, especially as it relates to the atoms, electrons, semiconductor materials and photons²¹ of modern devices, and

statistical mechanics, especially as it allows us to understand the thermal distributions of electrons and photons relevant for those devices, and to establish basic concepts like entropy.

The level here will be introductory, with the goal of understanding the key conceptual ideas so the reader can meaningfully reason about how such devices work.

We will necessarily be missing out other topics in modern physical science, not because they are unimportant, but because we need to concentrate on our goal. So, we will essentially be leaving out most of chemistry and the advanced quantum mechanics of elementary particles. We will also have to omit discussion of relativity. Relativity is by no means irrelevant for engineering – the finite and constant speed of light is very important in information transmission, and even general relativity is required if we are to engineer global positioning systems accurately, for example – but a sufficient introduction to it is beyond the scope we can cover here.

A brief summary of our intended topics is as follows.

²¹ Photons are the quantum mechanical particles that make up light.

Oscillations and waves

Ideas of oscillations and waves that we find in the classical world are important also as a background to quantum mechanics and to light, so we will introduce the description of oscillators and wave equations.

The quantum view of the world

The quantum view of the world is very different from the classical one. We will lead up to Schrödinger's wave equation and solve it for a simple example system so we see some of the key ideas in quantum mechanics.

Particles, atoms, and crystals

We will introduce key ideas of quantum mechanical particles, introduce the quantum mechanical model of the hydrogen atom, and discuss how we apply quantum mechanics to the crystalline materials of most modern electronic and optoelectronic devices.

Thermal distributions

The statistical interpretation of entropy allows us to understand concepts like the second law of thermodynamics and to deduce the forms for the thermal distributions of electrons and photons that underlie device operation and many other important physical processes.

Bands and electronic devices

We will use our understanding of the quantum mechanics and statistical mechanics of electrons in crystals to explain the basic principles behind electronic devices.

Light and quantum mechanics

We will look at the quantum mechanics of light and its absorption by and emission from materials.

Semiconductor optoelectronics

Using our understanding of light's quantum mechanics and the thermal distributions of electrons and photons, we will introduce the principles and concepts of optoelectronic devices in a wide range of applications, including communications, lighting, and solar energy.

Chapter 2

Oscillations and waves

2.1 Introduction

Before we start on new topics in modern physics, we review some basic classical physics that we need as background. Not surprisingly, before starting quantum mechanics, we need some ideas from classical mechanics. Though quantum mechanics is quite different from classical mechanics, it has to correspond back to it. Classical mechanics quite obviously describes much of the world around us quite effectively. So quantum mechanics should come to the same conclusions as classical mechanics in those situations. Hence, quantum mechanics in some sense will contain much of that classical mechanics¹.

There are also some key ideas from the classical world that are particularly important in setting up quantum mechanics. We need to be especially clear in our understanding of classical oscillating systems and waves. There is also a concept in the classical world for both oscillations and waves that is particularly important, both physically and mathematically, for setting up quantum mechanics. In the classical world, it is the idea of a “mode”, such as a mode of oscillation of some system. This is a term that, though heavily used in classical physics, is often not formally introduced as a general idea in discussions of classical mechanics. The mathematical idea of modes, which is essentially the same as the idea of what are called “eigenfunctions”, is one that is quite central to quantum mechanics.

So, here we introduce some of these basic ideas of classical mechanics, oscillators, modes, and waves. These are useful in both the classical and quantum worlds, and are essential background for quantum mechanics.

¹ We should state up front, though, that quantum mechanics cannot be “derived” from classical mechanics; quantum mechanics simply has to be postulated. Its ultimate justification is that it agrees with experiments, and you can be sure that it does agree very well with those experiments and that it indeed explains many things we did not previously understand.

2.2 Elementary classical mechanics

Momentum and kinetic energy

From elementary classical mechanics, we remember the idea of momentum. Momentum is the product of the mass m and the velocity $\mathbf{v}$.

$$\mathbf{p} = m\mathbf{v} \quad (2.1)$$

Mass is what we call a scalar – it is described simply by a number, here with “dimensions” of kilograms (kg) in the usual SI units. Velocity is a vector quantity that has both a magnitude² – a number, usually taken to be positive, and here in units of meters per second (m/s) – and a direction. As is common, we indicate vector quantities with a bold, non-italic font. Because momentum is a product of a number and a vector, it is a vector quantity as well.

The kinetic energy of a mass is the energy associated with its motion. From elementary classical mechanics, kinetic energy can be written

$$K.E. = \frac{1}{2}mv^2 \quad (2.2)$$

where v is the magnitude of the velocity. In more advanced treatments of classical mechanics, and as we make the transition towards quantum mechanics, it can be useful to write this in the equivalent form

$$K.E. = \frac{p^2}{2m} \quad (2.3)$$

where by p^2 we understand the vector dot product of $\mathbf{p}$ with itself, that is

$$p^2 = \mathbf{p} \cdot \mathbf{p} \quad (2.4)$$

We can understand v^2 in the same way as being $\mathbf{v} \cdot \mathbf{v}$.

Potential energy

Potential energy is defined as “energy due to position”. It is often denoted by the letter V . This choice of notation is particularly common in quantum mechanics. This V should not be confused with voltage, even if the potential energies we are working with are often electrostatic in origin. Often we call such potential energies just “potentials” for short. So this potential energy has units of joules, just like any other energy. (Voltage is the (electrostatic) potential energy per coulomb of charge, and so has units of joules per coulomb.)

² Often, we loosely refer to the magnitude of a vector as its “length”, though that “length” of a vector in this loose sense might not be a distance in ordinary space; the “length” of a velocity vector would be a quantity in units of meters per second, not meters, for example.

A key idea in potential energy is that an object will have the same potential energy at some point in space no matter what path it took to get there. Since we can write a position as a vector $\mathbf{r}$ relative to some position origin, we can write a potential³ as $V(\mathbf{r})$ – that is, we write potential energy as a function of position. Fig. 2.1 shows a potential “hill” or surface. The potential energy of some object at the point given by the dot is only a function of the position of the dot. Going round the closed path shown by the arrows back to where we started would make no difference to the potential energy.

Associated with some potential energy is a “field”. By field here we mean a force that is in general also a function of position, and we will discuss how to relate the two below. Classical “fields” with this property that “the energy associated with position depends only on the position and not how we got there” are called “conservative” or “irrotational”. Equivalently, the change in potential energy round any closed path is zero.

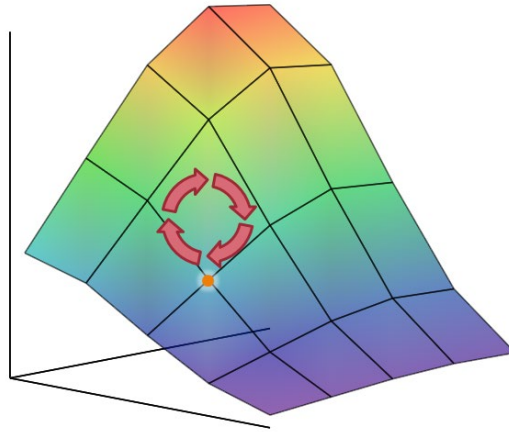


Fig. 2.1. Sketch of a potential surface. Larger potential energy corresponds to a larger value on the vertical axis.

Not all fields are conservative; for example, going round a vortex typically either requires energy to push you round it or you get energy out of it by being pushed round by it – in either case, you have a different energy when you get back to where you started. Various phenomena associated with changing magnetic fields are also non-conservative – you can change energy by going round a loop back to where you started. Gravitational and static electric fields are conservative, however.

Note, incidentally, that the “zero” or “origin” we use for potential energy is always arbitrary. We can choose it to be what we want as long as we are consistent. In both classical mechanics and quantum mechanics, there is no absolute origin for potential energy and we only really work with differences in potential energy between one

³ Note that, despite the confusion it generates, it is common, especially in quantum mechanics, to use the letter V to refer to potential energy, so a number of joules, even though we also commonly use V to refer to electrical voltage, which is joules/coulomb.

position and another. We do, however, typically make some specific choice of origin in practice, at least to make it easy to do the algebra in the problem of interest.

Force and potential energy

In classical mechanics, we often use the concept of force. Newton's Second Law gives us the relation between force $\mathbf{F}$ and acceleration $\mathbf{a}$ for a mass m

$$\mathbf{F} = m\mathbf{a} \quad (2.5)$$

Note that both the force $\mathbf{F}$ and the acceleration $\mathbf{a}$ are vector quantities. Acceleration is the rate of change of velocity. Equivalently to the statement in Eq. (2.5), we can write

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} \quad (2.6)$$

In other words, force is the rate of change of momentum.

When we have a potential energy, we can define the force corresponding to the potential energy as the “gradient” of the potential. To understand this idea, suppose we are trying to change the potential energy of a ball by pushing it slowly (and without friction) up a hill or a slope, as shown in Fig. 2.2.

We get a change ΔV in potential energy V of an amount

$$\Delta V = F_{push\ x} \Delta x \quad (2.7)$$

by exerting a force $F_{push\ x}$ through a distance Δx . Equivalently, we can say

$$F_{push\ x} = \frac{\Delta V}{\Delta x} \quad (2.8)$$

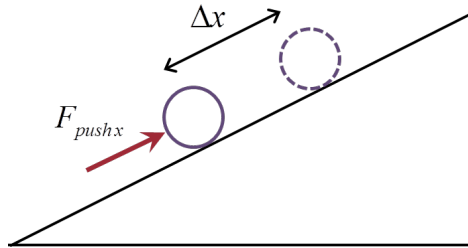


Fig. 2.2. A ball on an inclined plane, being pushed by a force of magnitude $F_{push\ x}$ through a distance Δx along the direction of the plane as shown.

In the limit as we take very small displacements, that is, as $\Delta x \rightarrow dx$ and $\Delta V \rightarrow dV$, we can write

$$F_{push\ x} = \frac{dV}{dx} \quad (2.9)$$

Now this is the “uphill” force we have to exert to push the ball up the “hill”. So the force F_x being exerted on the ball by the potential can be viewed as being a downhill force. So the relation between force and potential in this case is

$$F_x = -\frac{dV}{dx} \quad (2.10)$$

In general, the coordinate direction x will not necessarily line up like this along the direction of steepest change of the potential. For completeness, we can give the general relation between force and potential, with force again as a vector, and using the “gradient operator” ∇ , which we can define as we use it below,

$$\mathbf{F} = -\nabla V \equiv -\left[\frac{\partial V}{\partial x} \mathbf{i} + \frac{\partial V}{\partial y} \mathbf{j} + \frac{\partial V}{\partial z} \mathbf{k} \right] \quad (2.11)$$

where $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are unit vectors in the x , y , and z directions respectively. For example, for a mass at some position above the Earth’s surface, the gravitational force would be a vector pointing in the direction towards the center of the Earth; that force would technically be the gradient of the gravitational potential for that mass.

Quiz questions

- Q2.2.1 Which one(s) of the following correspond to vector quantities: speed, velocity, kinetic energy, potential energy, momentum?
- Q2.2.2 What is my kinetic energy if I weigh 60 kg and I am moving at 7 km/s: 2.1×10^2 J, 4.2×10^2 J, 1.47×10^3 J, 2.1×10^5 J, or 1.47×10^9 J?
- Q2.2.3 What are the correct units for the (magnitude of) momentum: J, kg m/s, m/s, $\text{kg}^2 \text{ m/s}$, or $\text{kg m}^2/\text{s}$?
- Q2.2.4 For the following question parts, recall that the magnitude of the force of gravity on an object with mass m near the Earth's surface is given by $F = mg$, $g \cong 10 \text{ m/s}^2$.
- a) Suppose I weigh 60 kg, and I walk up the stairs very slowly, gaining 2 m in height as I do so. What is the change in my potential energy as a result of climbing the stairs: 1200 J, 1.2 J, -1200 J, 0 J?
 - b) Suppose instead I run up the stairs at a speed of 0.5 m/s. What then is my change in potential energy as a result of climbing the stairs: 1.2 J, 1207.5 J, 1200 J, or 7.5 J?
 - c) Suppose I then walk 5 m along the corridor upstairs from the top of the stairs to get to my bedroom. What is my change in potential energy as a result of walking along the corridor: -3000 J, 3000 J, 1200 J, or 0 J?
 - d) Suppose I then walk up a second set of stairs, also 2 m in height, to the attic. What is the change in my potential energy as a result of climbing this second set of stairs: 2.4 J, 2400 J, 0 J or 1200 J?
 - e) Suppose I push my 70 kg sofa 3 m across a room, exerting a force of 200 N to do so. The sofa is not moving when I start to push it and is not moving when I stop. What is the change in the potential energy of the sofa as a result of me pushing it: 600 J, 2100 J, 0 J, or 2000 J?

Q2.2.5 Consider some smooth and flat plane that is 4 m long, and that is “inclined”, so it is higher at one end than the other. If the potential energy of a small ball is 10 J larger at the “high” end of the plane than at the “low” end of the plane, what is the force on the ball, in newtons (N), in the direction pointing down the plane (such a force would accelerate the ball, allowing it to roll down the plane): 2.5 N, 40 N, 160 N?

Problems

2.2.1*⁴ An electron with the charge $-e$ ($e = 1.602 \times 10^{-19} \text{ C}$ represents the elementary charge unit) is being accelerated by the electric field that is the gradient of some electric potential U , which is a function of position in the usual Cartesian coordinates x , y , and z . (Note: an electric potential can be the same thing as a voltage, and can be written in units of joules/coulomb (i.e., J/C).)

- a) Calculate the magnitude of the force driving an electron in the electric potential $U = (1x + 3y - 1.5z) \text{ J}/e$, where x , y , and z represent Cartesian coordinates.

[Notes: We are giving the electric potential energy in these slightly unusual units of joules/electron to make the calculation easier for you. The corresponding electric potential in volts would be $(1x + 3y - 1.5z)/e \text{ V}$ (or J/C). The potential energy in joules for an electron in this electric potential is simply $(1x + 3y - 1.5z) \text{ J}$. This question is an exercise in using the gradient operator as in Eq. 2.11. The force on this electron is simply the gradient of this potential energy expression $(1x + 3y - 1.5z)$. That result is a vector, and the magnitude of that vector is in newtons. Remember to use Pythagoras’s theorem to evaluate the magnitude of a vector. So, for some vector $\mathbf{f} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$, the “length” or magnitude of the vector is $f = \sqrt{a^2 + b^2 + c^2}$.]

- b) Which electric potential would impose a stronger magnitude of force on the electron, $U_1 = 2x$ or $U_2 = 1x + 1y$?

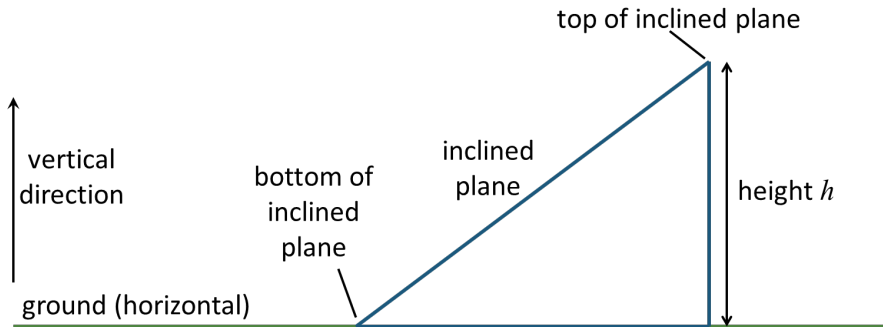
2.2.2 The force of gravity on a mass m is proportional to the mass. Near the surface of the Earth, it is a vertical force. Because of this proportionality, we can write the magnitude of this force as $F = mg$ where, near the surface of the Earth (the “ground”), the proportionality constant $g \approx 9.8 \text{ ms}^{-2}$ is typically referred to as the acceleration due to gravity. Here, we will consider the upward vertical direction to be positive.

- a) Suppose we have a mass $m = 2 \text{ kg}$ at some small distance above the ground. What is the force, in newtons, in the vertical direction on this mass due to gravity? (Be careful about the sign of your answer, and explain briefly why you choose whatever sign you choose.)

⁴ Note that solutions to all problems marked with an asterisk (*) are available in Appendix E.

- b) Suppose we choose the zero for potential energy to be at ground level, and that the mass m is 2 m above the ground. (To be more precise, we could say that the center of gravity of the mass is 2 m above the ground.) What is the potential energy of the mass m ? (Justify your answer briefly.)

Consider now the figure below of an inclined plane. The top of the plane is at a height h and the bottom of the plane is at a zero height.



- c) Suppose a small mass M is sitting right at the top of the inclined plane (initially motionless). Suppose also that we choose the zero for potential energy to be at ground level. Give an expression for the potential energy of the mass M , explaining briefly how you get this expression. [Hint: you can think of slowly lifting the mass up from the ground.]
- d) Suppose the mass now slides frictionlessly down the plane. Just before it gets to the very bottom of the plane,
- what is its kinetic energy? (Justify your answer briefly.) [Hint: think about conservation of energy.]
 - what is its potential energy? (Justify your answer briefly.)
 - what is its total energy (kinetic plus potential)?
- e) Suppose we now choose our zero of potential energy to be at the top of the inclined plane, and we again let the mass M slide frictionlessly down the plane. Just before it gets to the very bottom of the inclined plane
- what is its kinetic energy? (Justify your answer briefly.)
 - what is its potential energy? (Justify your answer briefly.)
 - what is its total energy (kinetic plus potential)?
- f) Continuing on from (e) (still choosing the top of the inclined plane as the zero of potential energy), suppose now the mass M has reached the bottom of the inclined plane, and after some time it comes to a stop (e.g., it might have slid along the ground somewhat, with friction bringing it to a stop) at ground level.
- what is its kinetic energy? (Justify your answer briefly.)
 - what is its potential energy? (Justify your answer briefly.)
 - what is its total energy (kinetic plus potential)?

2.3 Modes

Concepts and terminology

The concept of modes is one that appears in many aspects of physical systems. Despite its broad importance, it is often just introduced separately in each different area of physics rather than as a general idea that is useful across many fields. It is particularly important for us to introduce it here because this concept appears very often in quantum mechanics.

Modes occur in essentially all physical systems that oscillate, for example, or whenever we consider wave propagation. We find modes throughout classical physics, including applications in fields such as acoustics, mechanical structural engineering, and the electrical engineering of waves, signals, and oscillators. Modes can also have some quite remarkable and very useful mathematical properties. The concepts we find in the mathematics of modes are at the core of the mathematics of quantum mechanics. For all these reasons, it is well worth our time to spend some time finding out more about modes and some of their mathematics.

The terminology varies somewhat between different fields. In the classical physics of oscillations or waves, the term “mode” is common. In mechanical oscillations modes are also often known as *normal modes*. In the mathematics of modes, modes are more usually known as *eigenfunctions* or *basis functions*, and in quantum mechanics they are also known as *eigenstates* or *basis states*. In the case of quantum mechanics for atoms, they are also known as *orbitals*. The underlying concepts are the same in all these cases, however.

Examples of modes

The reader might ask “What is a mode?” Despite that fact that modes are very common in physics and engineering, it is difficult to find a broad definition of modes in text books. Most (and possibly all!) scientists and engineers who would insist that they know “perfectly well” what a mode is would be hard pressed to give a simple definition that was not rather mathematical. There is a precise mathematical answer, but it gives little direct physical insight, so we postpone it. Here we will discuss a few examples to get across the concepts of modes first.

We are quite used to the idea of modes in things that oscillate. In this case, the concept of a mode is essentially the same as that of a “resonance”, a preference to oscillate at a specific “resonant” frequency. Many such oscillators have only one way or mode in which they can oscillate (at least for small amplitudes of oscillation), and that way of oscillating is associated with only one frequency. Examples of this kind include a pendulum (at least if it is constrained to move in only one direction, as it usually is in a clock), and a mass on a spring (again if it is constrained to move in only one direction). You may have played a note on a bottle by blowing across the top; in this case, the oscillation is basically the air in the neck of the bottle behaving like a mass that is bouncing up and down on spring that is given by compressing the volume of air in the

body of the wine bottle. This kind of acoustic resonator is known as a Helmholtz resonator. The body of a guitar together with the hole in the front of it also forms a Helmholtz resonator⁵, and the Helmholtz resonator concept is extensively used in loudspeaker design, especially for boosting the bass response of loudspeakers. (It allows low frequency resonances even with resonator volumes that are much smaller than the wavelength of sound at such frequencies.)

Some other oscillating systems have many different resonances or modes. A guitar string has a fundamental mode, but also has harmonics that can also be excited, at twice the fundamental frequency, three times the fundamental frequency, and so on. The same is true for wind instruments such as the flute. Bodies that are essentially more two dimensional, such as a gong or a cymbal, have a much more complicated spectrum of resonances (which are in general not in integer ratios of frequencies), and three dimensional structures, such as a bell or a girder bridge, can have yet more complicated sets of possible resonances or modes. There are also electromagnetic resonators, basically metal cavities, that have various resonant modes, and there are optical resonant cavities of various kinds, formed with mirrors, that also similarly have multiple resonant modes.

In all these cases of oscillating modes or resonances, there are at least two common features. One is that each resonance or mode corresponds to quite a distinct way in which the object oscillates; the pattern of oscillation of the object is quite unique to a given mode, being quite different from the pattern of any other mode. As we will see later, the way in which the patterns of oscillation are different from one another can be quite specific mathematically; in a broad range of situations, the different patterns of oscillation are *orthogonal*, and we will be introducing that concept as we describe modes.

A second feature of oscillating modes is that, for each such resonance or mode, there is one well-defined numerical quantity, usually the frequency, associated with it. A third important point about all such modes is that, at least if we take a “loss-less” and “small amplitude” idealization of the mode, once excited, the oscillation would stay of exactly the same form forever, and everything that was oscillating would be oscillating at the same frequency in that mode.

Another kind of mode, called a “propagating mode”, arises in the propagation of waves, especially in waveguides. One notion essentially is that the wave stays the same shape as it moves. Essentially then, every point that propagates with the same “phase velocity”.

To proceed further in understanding modes, we need to analyze a few cases. That will begin to uncover the important mathematical properties and lead to a tighter definition of what a mode is.

⁵ In the case of the guitar, an effective “plug” of air in and around the hole behaves as the mass that bounces in and out against the “spring” that is the air inside the body of the guitar.

2.4 Analysis of systems with modes

Simple harmonic oscillator 2c

Perhaps the simplest system we can think of to illustrate modes is the “simple harmonic oscillator”. A good example is a mass on a spring, as shown in Fig. 2.3. We imagine the mass can only move up and down, not sideways, we presume it cannot twist or rotate in any way, and we presume that there is no friction.

The common dictionary definition of “mode” that is relevant here is that of a “manner” or “way”. There is only one “way” in which this mass can move, which is to oscillate up and down with a specific frequency.

The analysis of this system is straightforward. A simple spring will have a restoring force F in the y direction, acting on the mass M , and proportional to the amount y by which the spring is stretched. For some (positive) “spring constant” K , we can write that force⁶ as

$$F = -Ky \quad (2.12)$$

The minus sign is because this is a “restoring” force – it is trying to pull or push the position of the mass back towards its equilibrium value, which here we take to be at $y = 0$.

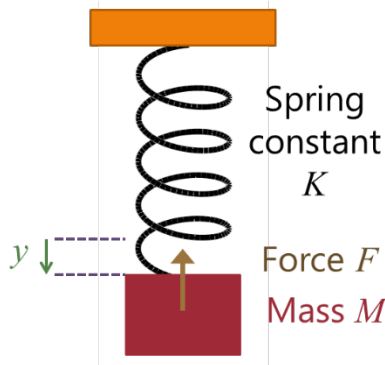


Fig. 2.3. A mass on a spring, without friction, oscillating up and down, here in the y direction.

The second equation that we need here comes from Newton’s Second Law, which gives us

$$F = Ma \quad (2.13)$$

where a is the acceleration of the mass in the y direction. The velocity v in the y direction is, by definition

⁶ Though quantities like force, velocity, and acceleration are all generally vectors, because we are only dealing with motion in one direction – here the y direction – we can treat them effectively like scalars, assuming the y vector direction in all cases.

$$v = \frac{dy}{dt} \quad (2.14)$$

which is the rate of change of position, y , with respect to time, t . Acceleration is the rate of change of velocity, so here

$$a \equiv \frac{dv}{dt} \equiv \frac{d}{dt} \left(\frac{dy}{dt} \right) \equiv \frac{d^2 y}{dt^2} \quad (2.15)$$

So, substituting in Eq. (2.13) gives

$$F = Ma = M \frac{d^2 y}{dt^2} \quad (2.16)$$

Substituting using Eq. (2.12) for F gives

$$F = Ma = M \frac{d^2 y}{dt^2} = -Ky \quad (2.17)$$

Defining

$$\omega^2 = K / M \quad (2.18)$$

or, equivalently,

$$\omega = \sqrt{K / M} \quad (2.19)$$

where we take the positive square root, and rearranging Eq. (2.17) gives

$$\frac{d^2 y}{dt^2} = -\frac{K}{M} y = -\omega^2 y \quad (2.20)$$

Mathematically, one solution to an equation like this is

$$y = \sin \omega t \quad (2.21)$$

which is easy to check by substitution. Explicitly

$$\frac{d}{dt} \sin \omega t = \omega \cos \omega t \quad (2.22)$$

and

$$\frac{d^2}{dt^2} \sin \omega t = \frac{d}{dt} \omega \cos \omega t = \omega \frac{d}{dt} \cos \omega t = \omega [-\omega \sin \omega t] = -\omega^2 \sin \omega t \quad (2.23)$$

This sine solution shows that this equation describes oscillations. The sine function is periodic in its argument, repeating every 2π . So for some amount of time Δt such that

$$\omega \Delta t = 2\pi \quad (2.24)$$

the sine function will be back to where it started. Hence Δt is what we can call the “period” of the oscillation. Quite generally, the frequency of an oscillation, in cycles

per second or hertz (Hz), is the inverse of the period; that is, this oscillation has a frequency

$$f = \frac{1}{\Delta t} \quad (2.25)$$

so the relation between the frequency f and ω is

$$\omega = 2\pi f \quad (2.26)$$

In physics, it is very common to use this quantity ω , which is called the “angular frequency” rather than the (conventional) frequency f , mostly just because it saves writing down “ 2π ” so often. We can also think of angular frequency as the number of radians per second. There are 2π radians in one “circle” or cycle here, so the angular frequency is 2π times larger than the (conventional) frequency.

For Eq. (2.20), we can easily check that $y = A \sin(\omega t)$ would be a solution, where A is any constant, which just says that, at least in our idealized mathematical model here, any amplitude (here A) of oscillation is possible. Also $y = B \cos \omega t$ would be a solution for any constant B . We could write the most general solution of Eq. (2.20) as

$$y = A \sin \omega t + B \cos \omega t \quad (2.27)$$

As is typical for such “second order” equations (that is, in which the highest order of derivative involved is the second derivative), we have two arbitrary constants, here A and B , in this general solution. Just what the values of those are in any particular situation depends on the starting conditions – basically, when we started the oscillation and how big the amplitude of it was.

The only real differences between the solution in Eq. (2.21) and the general solution Eq. (2.27) are that the general solution of Eq. (2.27) allows for any amplitude and any “phase” of oscillation. (The sine and cosine functions are 90 degrees out of phase with one another, so combinations of them can represent an oscillation of any phase.) No matter which solution we choose here, the mass oscillates up and down at angular frequency ω .

Other than having different possible amplitudes and phases of oscillation, there is only one way in which this simple harmonic oscillator can oscillate. It has only one mode of oscillation.

The concept of modes is most useful for linear systems, that is, essentially, ones whose behavior is the same regardless of the amplitude of the oscillation. If some function $y(t)$ is a solution of Eq. (2.21), so also is $C y(t)$ where C is any constant. This is obvious from the general solution as in Eq. (2.27) – multiplying such a solution just corresponds to replacing A and B with new constant numbers – and it is easily verified directly by substituting in Eq. (2.21).

Quiz question

Q2.4.1 Suppose I have a “mass on a spring” oscillator that oscillates at 100 Hz. What, approximately, is the angular frequency corresponding to this oscillation in radians per second: 628, 15.9, 100, or 157?

Eigen equations

Note that the equation Eq. (2.20) of our linear oscillator is of the form

$$\text{Operator operating on the function } y(t) = \text{a constant times the function } y(t) \quad (2.28)$$

where here the operator⁷ is d^2 / dt^2 and the constant is $-\omega^2 = -K / M$

If we denote an operator by A and functions by y and z respectively, then we can introduce a mathematical notation, and write

$$Ay = z \quad (2.29)$$

where we mean that operator A operating on function y gives rise to function z . With that notation, we can write Eq. (2.28) in general as

$$Ay = by \quad (2.30)$$

where A represents the operator, y represents the function, and b represents a constant.

The form of equation given by relation (2.28) (or (2.30)) is called an eigen equation⁸. Any constant b for which such an equation holds is called an eigenvalue, and a function that is a solution for that eigenvalue is called an eigenfunction associated with that eigenvalue⁹. The particular eigen equation Eq. (2.20) for our physical problem of a mass on a spring, for a given K and M , physically is only allowed to have one eigenvalue ($-K / M$) and one eigenfunction (at least if we allow different amplitudes and phases). As we stated above, this particular oscillator is an example of an oscillator that physically only has one mode¹⁰.

⁷ An “operator” is something that turns one function into another function.

⁸ “eigen” is a German word with several possible translations into English depending in the context; one translation is “own” as in the sense of “one’s own”.

⁹ It is possible in general to have more than one eigenfunction associated with a given eigenvalue, in which case the eigenvalue is said to be degenerate. Incidentally, for the equations associated with linear physical problems, there can only be a finite number of such eigenfunctions associated with a given eigenvalue; that number is called the degeneracy of the eigenvalue.

¹⁰ One subtlety here is whether we regard the sine and cosine version of this oscillation as being different modes. It is possible to take that approach of considering these as different modes, but it is more common to say that there is one mode with two attributes associated with it, one being amplitude and the other being phase.

Linear operators

An important point about this mathematics we are discussing here is that the operator d^2 / dt^2 is what is called a linear operator. The definition of a linear operator is that it should obey the relations

$$(i) \quad \text{Operator operating on } (b \text{ times } y) = b \text{ times (Operator operating on } y) \quad (2.31)$$

for an arbitrary constant b and a function y , and

$$(ii) \quad \text{Operator operating on } (y_1 + y_2) = \\ \text{(Operator operating on } y_1) + \text{(Operator operating on } y_2) \quad (2.32)$$

for any two functions y_1 and y_2 .

By this definition, it is straightforward to see that d^2 / dt^2 is linear.

In our more abstract operator algebra, incidentally, we could write Eq. (2.31) as

$$Ab\mathbf{x} = bA\mathbf{x} \quad (2.33)$$

and Eq. (2.32) as

$$A(\mathbf{x}_1 + \mathbf{x}_2) = A\mathbf{x}_1 + A\mathbf{x}_2 \quad (2.34)$$

If the notation of Eqs. (2.29), (2.30), (2.33) and (2.34) seems somewhat abstract, note that for such linear operators it is exactly the same as matrix and vector notation. We can think of A as a matrix¹¹, and each of the $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ objects as being column vectors¹².

By now the reader can see that we are moving towards defining modes as the eigenfunctions of an operator, especially a linear operator, and that is indeed the case.

Quiz questions

Q2.4.2 The mass-on-a-spring oscillator has an eigen equation that can be written in the form $\frac{d^2 y}{dt^2} = -\lambda y$. If the frequency of oscillation is 10 Hz, what, approximately, is the value of λ ? 10 Hz, 10^2 Hz^2 , $2\pi \times 10$ radians per second, or $4\pi^2 \times 10^2$ (radians per second)²?

¹¹ Quite generally, linear operators can be represented by matrices, though we will postpone any general discussion of that equivalence.

¹² Here we mean vectors in the general matrix-vector sense of vectors as columns of numbers. These are not necessarily geometrical vectors, though they could be if we were considering column vectors with three elements. Here $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ just arbitrary mathematical vectors; we are not presuming they have any specific relation to x , y , and z directions in Cartesian geometrical space, nor even that they are vectors in three or any other specific number of dimensions.

Q2.4.3 Suppose A is a linear operator that operates on functions $\mathbf{x}$. Choose whether each of the following statements is necessarily true. (For the statement to be necessarily true, it would have to be true for all possible choices of $\mathbf{x}$ and/or $\mathbf{y}$).

- a) $A(\mathbf{x} + \mathbf{x}) = A\mathbf{x} + A\mathbf{x}$
- b) $A(3\mathbf{x}) = A\mathbf{x} + 3$
- c) $A(\mathbf{x} + \mathbf{y}) = A\mathbf{x} + \mathbf{y}$
- d) $A(\mathbf{x} + 3\mathbf{y}) = A\mathbf{x} + 3A\mathbf{y}$

Problems

2.4.1* For a harmonic oscillator consisting of a mass M on a spring of spring constant K , which will have an oscillation (angular) frequency of $\omega = \sqrt{K/M}$, what are the “dimensions” of the spring constant K ?

By “dimensions” we mean what unit or combination of units from the SI system of units describes the units in which we write K . SI base units include kg (kilograms), m (meters), s (seconds), and A (amperes) as well as other derived units. Express your answer in two forms:

- (i) using only “base” units (i.e., kg, m, s, A)
- (ii) using N (newtons) (the “derived unit” of force) as one of the units in your answer.

2.4.2 If we have a simple harmonic oscillator with a mass of 1 g (i.e., one gram) that will oscillate at a frequency of 1 kHz (note – this is the conventional frequency, not the angular frequency), what is the value of the spring constant K expressed in the appropriate SI units?

2.4.3 For the simple harmonic oscillator of question 2.4.2, if we start the oscillator out by pulling the mass by a distance of 1 mm along the oscillating direction (e.g., the y direction), thereby stretching the spring by 1 mm, how much potential energy do we initially put into the system? (Hint: remember that we can deduce the change in potential energy by integrating the product of force times distance. The distance starts out at $y = 0$ (which is the “equilibrium” position the mass will sit at if it is not oscillating) and ends up at $y = 1$ mm, and the force is proportional to the amount by which the spring is stretched relative to zero.)

2.4.4 For this same simple harmonic oscillator as in questions 2.4.2 and 2.4.3, once we have released it from its starting position as in question 2.4.3 above, how much kinetic energy does the mass have when it reaches $y = 0$? (Hint: if you think about it, you don’t have to calculate anything here!)

2.4.5* A box of 2 kg is attached to a spring with constant $K = 10$ N/m and moves in the y direction, stretching or compressing the spring as it does so. If the box is displaced to a point $y = 2$ m relative to its equilibrium point and then released, how long does it take to travel back to $y = 1$ m? What then is the speed of the mass at this point $y = 1$ m?

- 2.4.6 Suppose we consider a simple harmonic oscillator whose oscillation we have written down as

$$y(t) = D \cos(\omega t + \phi)$$

where D , ω and ϕ are constants. (This is often quite a convenient form. We could then call D the amplitude of the oscillation, and ϕ the phase.) Show that we can also write this in the form

$$y(t) = A \sin \omega t + B \cos \omega t$$

and give expressions for A and B in terms of D and ϕ .

- 2.4.7 A simple pendulum is one that can be considered to be a point mass suspended from a string of negligible mass and of length L . For small maximum angular deviation $\theta_{\max}$ from the vertical direction, the motion of the point mass in angle as a function of time is very similar to the motion of the position as a function of time of a mass on a spring. Following similar procedures to those used above to derive the oscillation frequency of a mass on a spring, derive an expression for the oscillation frequency of this simple pendulum. (Hints: You might find you want to write an expression for the potential energy of the mass as a function of angle θ , or alternatively you might evaluate an appropriate force on the mass as a function of angle θ , taking some small angle approximation to an appropriate trigonometric function.) You may take the acceleration due to gravity to be a constant g . (You do not need to evaluate your answer numerically – an appropriate algebraic expression will be sufficient.)

- 2.4.8 Consider the equation

$$\frac{dy}{dt} = iby$$

where b is a real number that we can choose to be any (non-zero) value and $i = \sqrt{-1}$. This equation is written in the form of an eigen equation because it has an operator, d/dt , acting on a function y (which we could also write explicitly as a function $y(t)$) on the left, and on the right it has a number (here the imaginary number ib) times that same function. In general, this number ib is the eigenvalue in such an equation.

For each of the following functions, state whether it is an eigenfunction of this equation and, if so, state the corresponding eigenvalue. [For example, the function $y(t) = \exp(3it)$ is an eigenfunction with eigenvalue $3i$ (so $b = 3$; this

is a real number as required) because $\frac{dy}{dt} = \frac{d}{dt} \exp(3it) = 3i \exp(3it) = 3iy$]

- a) $y = 4 \exp(3it)$
- b) $y = \exp(-3it)$
- c) $y = \exp(3it) + \exp(-3it)$

d) $y = \exp(4it + 3.5)$

e) $y = \exp(2t^2)$

f) $y = 5t$

g) $y = \exp(1.5t)$

2.4.9 An operator A acting on a variable y , which we will write here using the notation Ay , is linear if and only if it meets the two criteria for linear operators (Eqs. (2.31) and (2.32)) for all possible values of y (and we take y to be any real number). For the operators listed below, identify which are linear operators, explaining your answers in terms of these two criteria. (You may think of variables x and y as being different position variables corresponding to different coordinate axes, and variable t as corresponding to time, though these correspondences are not actually necessary to address this question.)

a) $Ay \equiv y^2$

b) $Ay \equiv \frac{dy}{dx} - e^x y$

c) $Ay \equiv \frac{dy}{dt} + \sin(y)$

d) $Ay \equiv \frac{d^2 y}{dt^2} - y - 2$

2.5 The classical wave equation

Waves are very common in classical physics. For example, we encounter sound waves through air, waves on the surface of water, and electromagnetic waves that we exploit for radio and light. We will also find other kinds of waves in quantum mechanics. Here, by looking at waves on a string, we introduce a simple example of a classical wave equation. Once we continue in the next section to look at standing waves on a string, we will see that this wave equation can also have various different modes associated with it.

We imagine first that we have a set of small, identical masses that are connected in a horizontal line by a string. We pretend that the string itself has no mass. There is, however, a “tension” T in the string. We can imagine, if we like, that the far ends of the string are tied to posts, and we have tightened the string between the posts to give it this tension.

The meaning of the tension is that there is a force of magnitude T pulling on each of these posts. If we were to imagine cutting the string at some point in the middle, and inserting ourselves in between the two lengths of string, to retain the tension in the string, we would also have to pull each length of string, one with our left hand and the other with our right hand, each with a force T . This tension therefore extends throughout the string. If the string is straight, then each mass has equal and opposite forces pulling

it horizontally from both sides, just like the forces on our left and right hands when we insert ourselves in the middle of the string.

Consider, then, such masses separated by individual short lengths of string, each Δz in length. For simplicity, we presume the masses are actually arbitrarily “short” in the z direction along the string¹³, so the center-to-center spacing of the masses can also be taken as Δz . Suppose, though, at some instant the masses are each possibly displaced vertically, as shown in Fig. 2.4. The y coordinates show the amount of displacement of each mass, and we look specifically at the j th such mass.

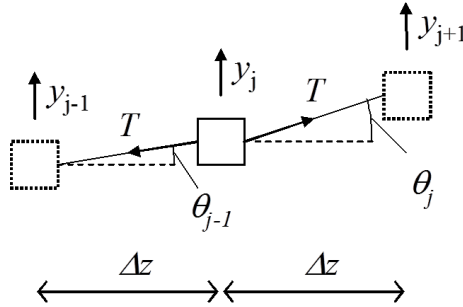


Fig. 2.4. Masses connected by a string of tension T .

We want to know the net force in the y direction on the j th mass. A force $T \sin \theta_j$ pulls mass j upwards from the right hand side. This is the resolved component of force acting in the vertical direction, from the tension in the string on the right of the mass. Similarly, a force $T \sin \theta_{j-1}$ pulls mass j downwards from the left hand side. So the net upwards force on mass j is

$$F_j = T(\sin \theta_j - \sin \theta_{j-1}) \quad (2.35)$$

The vertical displacements and angles are exaggerated in Fig. 2.4, and we presume that all the angles are actually very small. As a result, the sine can be approximated by the tangent, so

$$\sin \theta_j \simeq \frac{y_{j+1} - y_j}{\Delta z} \quad \text{and} \quad \sin \theta_{j-1} \simeq \frac{y_j - y_{j-1}}{\Delta z} \quad (2.36)$$

So Eq. (2.35) becomes, approximately

$$F_j \simeq T \left[\frac{y_{j+1} - y_j}{\Delta z} - \left(\frac{y_j - y_{j-1}}{\Delta z} \right) \right] = T \left[\frac{y_{j+1} - 2y_j + y_{j-1}}{\Delta z} \right] \quad (2.37)$$

We can multiply the top and bottom lines in Eq. (2.37) by Δz to obtain

$$F_j = T \Delta z \left[\frac{y_{j+1} - 2y_j + y_{j-1}}{(\Delta z)^2} \right] \quad (2.38)$$

¹³ We have to draw them with some finite length, of course, in the figure.

But the expression in square brackets, in the limit of small Δz , is just the second derivative of the coordinate y with respect to the horizontal coordinate z . So we can write, in the limit of small Δz , that the force on the mass j is

$$F = T \Delta z \frac{\partial^2 y}{\partial z^2} \quad (2.39)$$

(We are writing the second *partial* derivative with respect to position z here because we are only interested in the result at a specific time, so technically this is a second derivative in space at a particular time. We can also regard this partial derivative as the second derivative in space of a “snapshot” of the string at some specific time.)

Note that, with Eq. (2.39), we are saying that the vertical force F on the mass at a given point is proportional to the curvature of the “string” of masses. There is no net vertical force if the masses are in a straight line, even if that line is “tilted” at an angle. We have to have a difference in the slopes or “tilting” of the pieces of string on either side of a mass before there is any net force on the mass.

Now we can think of the masses in Fig. 2.4 as the amount of mass per unit length in the z direction, that is, the linear mass density ρ (that is, kilograms per meter), times Δz , which we can write as a mass $m = \rho \Delta z$. We can also write from Newton’s Second Law for that element of mass that

$$F = m \frac{\partial^2 y}{\partial t^2} = \rho \Delta z \frac{\partial^2 y}{\partial t^2} \quad (2.40)$$

(Now we are taking the second *partial* derivative with respect to time here because we are interested in the behavior in time of an element of mass at a particular point in space.) Putting Eq. (2.39) and Eq. (2.40) together and cancelling the Δz on both sides gives

$$\frac{\partial^2 y}{\partial z^2} = \frac{\rho}{T} \frac{\partial^2 y}{\partial t^2} \quad (2.41)$$

Rewriting Eq. (2.41) with

$$v^2 = T / \rho \quad (2.42)$$

gives

$$\frac{\partial^2 y}{\partial z^2} - \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2} = 0 \quad (2.43)$$

Though we have not justified the meaning of v yet, it will correspond to wave propagation velocity, and Eq. (2.43) is a wave equation for a wave with velocity v . We can write

$$v = \sqrt{T / \rho} \quad (2.44)$$

where we take the positive square root because we like to talk conventionally about a positive wave velocity¹⁴. Hence we have constructed a wave equation, Eq. (2.43) for waves on a string.

We can check by direct substitution into the wave equation Eq. (2.43) that any function of the form $f(z - vt)$ is a possible solution of such an equation. This corresponds to a “pulse” propagating to the “right”. Similarly, any function of the form $g(z + vt)$ is also a possible solution, corresponding to a “pulse” propagating to the “left”. Furthermore, any sum of such solutions is also a solution.

Quiz questions

Q2.5.1 A uniform string under a tension of 10 N has a wave velocity of 500 m/s. If I increase the tension to 90 N, what now is the wave velocity?

- a) It increases by 9
- b) It decreases by 9
- c) It increases by 3
- d) It decreases by 3

Q2.5.2 Suppose we have a string of uniform linear density and under some specific tension. If some function $h(z - vt)$ is a possible wave on this string, state for each of the following whether it is necessarily a possible wave on the string also.

- a) $-h(z - vt)$
- b) $h(z) - vh(t)$
- c) $h(z + vt)$

Problems

2.5.1 Is the following statement true or false? (Justify your answer briefly.)

“In the classical wave equation $\frac{\partial^2 y}{\partial z^2} - \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2} = 0$ for a nominally horizontal string stretched in the z direction, where y is the displacement of the string in the vertical direction, the acceleration of the string in the vertical direction at a given point is proportional to the curvature of the string at that point (as given by the

¹⁴ In this equation we could also take the negative square root, and in fact both the positive and negative results have meaning – they correspond to waves propagating in positive (e.g., to the “right” and negative (e.g., to the “left”) directions. Conventionally, though, we use the positive value for v , and use positive or negative signs in front of it to describe waves in the two directions. Of course, velocity is more correctly a vector, but typically nonetheless we refer to this positive magnitude of the velocity as just the wave propagation velocity, with “magnitude” often just being understood by context.

second derivative of the string displacement) (with a positive proportionality constant).”

- 2.5.2 Consider an infinitely long string for which the wave propagation velocity is v . (The infinite length means that there are no reflections off any “ends” of the string, and any wave that satisfies the general solution of the wave equation for the string is a valid solution). For each of the following proposed functions, which represent the y displacement of the string, state whether it can be a solution of the wave equation for this string. (The string is presumed to be along the “horizontal” z direction, and y corresponds to a “vertical” displacement of the string.) In the following functions we presume A and B are constants, and, where both k and ω are present in the expression, we presume that $k = \omega / v$. In all cases, k is a constant. [Note: In most cases, it should be possible to deduce the answers here without substituting directly into the wave equation to check if these are solutions, though you can always check these that way also.]

- (i) $A \sin(kz - \omega t)$
- (ii) $A \sin(kz - \omega t) + B \cos(kz + \omega t)$
- (iii) $A \exp(-k^2 z^2)$
- (iv) $A \cos(\omega t) \sin(kz)$
- (v) $A \cos(\omega t) \exp(-k^2 z^2)$
- (vi) $A \exp[-(kz - \omega t)^2]$
- (vii) $A \sin(kz - \omega t) + B \sin(2kz - 2\omega t)$

2.6 The Helmholtz equation

Often, we are interested in waves that are oscillations at one specific (angular) frequency ω – what we can call “monochromatic” waves¹⁵. The time dependence of such waves could be of the form $\sin \omega t$ or $\cos \omega t$ or some combination of the two. In that case, the second partial derivative with respect to time of some such function $y(z, t)$ always has the form

$$\frac{\partial^2 y}{\partial t^2} = -\omega^2 y \quad (2.45)$$

Since we presume we know that the form of y as a function of t is just oscillatory at (angular) frequency ω , we can now simplify our notation for this monochromatic case. From now on, we only need to consider how y varies with z , so from now on we need only consider y as a function of z ; we can always go back and include the time dependence by multiplying by an oscillating function at angular frequency ω . So, using this result Eq. (2.45), we rewrite our wave equation Eq. (2.43) as

¹⁵ “Monochromatic” literally means “one color”.

$$\frac{d^2 y}{dz^2} + k^2 y = 0 \quad (2.46)$$

or, if we prefer,

$$\frac{d^2 y}{dz^2} = -k^2 y \quad (2.47)$$

where

$$k^2 = \frac{\omega^2}{v^2} \quad (2.48)$$

Equivalently, we can imagine we are taking a snapshot of this monochromatic wave in space at some time, and that snapshot now has to obey the Eq. (2.46). Note now that, since we are regarding y as only being a function of the one variable z , we can write the conventional “total” or “ordinary” derivative again (rather than the “partial” derivatives we needed in Eq. (2.43) when we had to include all possibilities associated with y being a function of both z and t).

The quantity k in Eq. (2.48) is often called the “magnitude of the wavevector” or, where there is no confusion about directions, just the wavevector¹⁶. Note, incidentally, that in discussing waves mathematically, it is often more convenient and generally more common to use this idea of a wavevector magnitude rather than the wavelength. Quite generally, if a monochromatic wave has a wavelength λ , then, by definition, it has a wavevector magnitude

$$k = \frac{2\pi}{\lambda} \quad (2.49)$$

A wavelength λ is the distance to go through a complete cycle in a monochromatic wave. The quantity $1/\lambda$ would be the number of wavelengths per unit length (e.g., the number of wavelengths per meter). The quantity $k = 2\pi/\lambda$ is then the number of radians per unit length (e.g., radians per meter), there being 2π radians in one complete cycle. A wave in space in the z direction with a wavelength might have the form $\sin(2\pi z/\lambda)$; such a form would correctly repeat in z every wavelength. Using the wavevector magnitude, k , we could write this more compactly as $\sin(kz)$. Here, since we will mostly just discuss waves in one direction for simplicity, it may seem odd to refer to this quantity k as the magnitude of some vector. More generally, and very usefully when we are working with waves in three dimensions, we can write the vector quantity $\mathbf{k}$, which has the magnitude k and the vector direction corresponding to the direction in which the wave is propagating.

¹⁶ Again, not surprisingly given the name “wavevector”, this “wavevector” is really more correctly a vector, not a scalar. However, there is no other good word for this k , and hence it is typically just called the wavevector (or wavevector magnitude), with the understanding that we will write it more correctly as a vector when it matters. For waves in one direction, we just write it and name it in this scalar fashion.

A wave equation in the form of Eq. (2.46) is called a Helmholtz wave equation, and it is essentially the simplest wave equation we can write down that describes the behavior in space of a monochromatic wave. Mathematically, we recognize that the Helmholtz equation Eq. (2.47) is also in the form of an eigen equation, just like Eq. (2.20) for the harmonic oscillator. The eigenvalue now is $-k^2$, though the Helmholtz equation is a differential equation in z , not t . Just as before for the simple harmonic oscillator, this equation can be written in an abstract mathematical form as

$$A\mathbf{y} = -k^2\mathbf{y} \quad (2.50)$$

where now $\mathbf{y}$ represents the entire function $y(z)$ – for our string example, the “snapshot” of the string at some time when there is a wave on it oscillating at an angular frequency ω – and A corresponds now to the linear operator d^2/dz^2 .

When we were considering the simple harmonic oscillator, we had an equation in the same overall mathematical form, that is, Eq. (2.20), but the eigenvalue was fixed for us by the choice of the spring constant and the mass. In the case of the Helmholtz equation for a string, for example, we may have chosen a definite density ρ and a definite tension T , but those only set the magnitude v of the wave velocity. On its own, as we see from Eqs. (2.42) and (2.48), this wave velocity v only determines the ratio of ω^2 and k^2 , that is,

$$v^2 = \frac{T}{\rho} = \frac{\omega^2}{k^2} \quad (2.51)$$

For this eigen equation Eq. (2.50), unless we impose some other constraints on the problem, any eigenvalue $-k^2$ is possible, which in turn means any frequency is possible for our monochromatic wave. This should not surprise us physically; we can imagine a wave of any particular frequency moving down a string.

Finally here, we note that we can generalize the Helmholtz equation from the “one-dimensional” form of Eq. (2.46) to a “three-dimensional form”, which is

$$\nabla^2\psi(\mathbf{r}) + k^2\psi(\mathbf{r}) = 0 \quad (2.52)$$

where

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \quad (2.53)$$

Here we are using the Greek letter ψ to represent the wave amplitude to avoid any confusion between the coordinate directions x , y , and z . We have written the wave amplitude as a function of the vector position $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ where $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are unit vectors in the x , y , and z directions¹⁷.

¹⁷ The different uses of the letter “ k ”, as a wavevector magnitude and, in bold form “ $\mathbf{k}$ ” as both the unit vector in the z direction and as a wavevector, is so common in the physics of waves that we will also use “ k ” in these different forms here. The reader will unfortunately just have to get used to this confusing notation both here and in the rest of the physics literature, though we will try to be clear here whenever there is actual possibility of confusion.

In Eq. (2.52), we need to be using partial derivatives because we have three different spatial directions. So $\partial^2 / \partial x^2$ to the second derivative with respect to x at specific values of y and z coordinates. Of course, we are also still imagining that this whole Helmholtz equation refers to a “snapshot” in time of a wave that is oscillating at (angular) frequency ω .

The three-dimensional Helmholtz equation Eq. (2.52) is useful, for example, for modeling acoustic waves in air. We will also be using it later as we start quantum mechanics, though we will mostly be using the simpler one-dimensional version, Eq. (2.46).

Quiz questions

Q2.6.1 Which of these, suitably rearranged, is a Helmholtz wave equation, for some appropriate positive real “wavevector magnitude” k ? (Choose all correct answers (and no wrong ones!))

a) $\frac{d^2 y(z)}{dz^2} = -k^2 y(z)$

b) $\frac{-1}{k^2} \frac{d^2 y(z)}{dz^2} = y(z)$

c) $\frac{d^2 y(z)}{dt^2} + k^2 y(z) = 0$

Q2.6.2 When starting from the full wave equation for a given uniform string under a specific tension, we can choose any value of (angular) frequency ω when constructing a Helmholtz equation from it. True or False?

Q2.6.3 In considering the Helmholtz equation for a uniform string under tension, if we double the (angular) frequency ω , the value of the corresponding value of k will change by which of the following amounts: halve, double, or increase by a factor of 4?

Problems

2.6.1 Is the following statement true or false? “The Helmholtz equation for a wave on a string tells us that, for a monochromatic wave, the curvature of the string at a given point (as given by the second derivative of the string displacement) is always proportional to the displacement of the string (with a constant but negative proportionality constant) at that point.”

2.6.2 For the functions listed below, identify which are solutions to the Helmholtz wave equation for some specific magnitude of the wave velocity in each case. (We are presuming a “string” with no particular boundary conditions at the end, which we could view as an infinitely long string, i.e., there are no boundary conditions to constrain possible solutions here, though we are presuming some specific tension and linear mass density for the string in each case.)

$$(a) \quad y(z, t) = \sin\left(\frac{z-3t}{2}\right) + \sin\left(\frac{z+3t}{4}\right)$$

$$(b) \quad y(z, t) = \sin\left(\frac{z}{2}\right)\sin(2t) + \sin\left(\frac{z}{4}\right)\sin(2t)$$

$$(c) \quad y(z, t) = \sin\left(\frac{z}{2}\right)\sin(2t) + \cos\left(\frac{z}{2}\right)\sin(4t)$$

2.7 Standing waves 3d

So far, we have put no constraints on the solutions of the (one-dimensional) Helmholtz equation, Eq. (2.46). Just as we found the general solution Eq. (2.27) for the simple harmonic oscillator equation Eq. (2.20), we similarly have a general solution for this Helmholtz equation Eq. (2.46), because these equations have the same mathematical form. So our general solution here can be written

$$y = A \sin kz + B \cos kz \quad (2.54)$$

where A and B can be any constants; any such wave could be a solution at the moment.

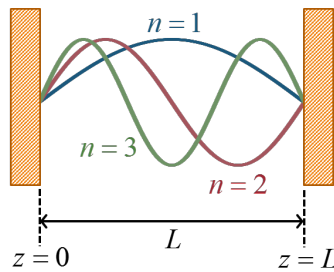


Fig. 2.5. Standing waves on a string between two walls a distance L apart.

Now suppose that the string is tied to rigid posts or walls at either end, as in Fig. 2.5. Now we are imposing boundary conditions. For posts or walls a distance L apart, we require that the net wave is zero at both of those (rigid) points. We can choose one of these posts or walls to be at $z=0$, and so the other is at $z=L$. So, we require

$$y(z) = 0 \text{ for } z = 0, L \quad (2.55)$$

Now we know that, at $z=0$, $\sin kz \equiv \sin 0 = 0$ and $\cos kz \equiv \cos 0 = 1$. So to have $y(0) = 0$ in a solution of the form of Eq. (2.54) we require $B = 0$. Hence, given our boundary conditions, solutions of this equation are of the form

$$y(z) = A \sin kz \quad (2.56)$$

for some k and for an arbitrary constant A . We also require that $y(L) = 0$ (i.e., at the right hand end of the string). Now, $\sin \theta = 0$ requires that $\theta = n\pi$ where n is an integer. So we therefore require that k takes one of the values

$$k_n = n\pi / L \quad (2.57)$$

where n is an integer. These, then, give the possible eigenvalues for this problem¹⁸. Hence the resulting eigenfunctions are

$$y_n(z) = A \sin\left(\frac{n\pi z}{L}\right) \quad (2.58)$$

The case of $n = 0$ is a trivial one, because the wave would then be zero everywhere on the line. Also, solutions for negative n are really the same solutions as for positive n . (Since $\sin \theta = -\sin(-\theta)$, changing from n to $-n$ just corresponds to replacing A with $-A$; since A is an arbitrary number anyway, there is no real difference between the solutions for positive and negative values of n .) Hence, we only need to consider $n = 1, 2, 3, \dots$.

These solutions, Eq. (2.58), are simple standing waves on the string. These different standing waves are the (eigen) modes of this problem. The first three of these are illustrated in Fig. 2.5.

For a given string with a given density and tension, we know there is a specific relation, Eq. (2.48), between the wavevector magnitude k and the (angular) frequency ω . Since we can choose both of those to be positive numbers, we can rewrite Eq. (2.48) as

$$\omega = vk \quad (2.59)$$

So, because we have concluded there are only specific values of k allowed given our boundary conditions here – explicitly those given by Eq.(2.57) – then we similarly only have a corresponding set of allowed values of the (angular) frequency

$$\omega_n = vk_n = \frac{n\pi v}{L} \quad (2.60)$$

So, with a given string stretched with a given tension (which together lead to a specific wave velocity magnitude v), and anchored at two posts or walls, there are only specific frequencies of oscillation possible. In other words, insisting that we are looking only for solutions that correspond to monochromatic waves, in such a situation with the string tied between two rigid posts or walls, such solutions or “oscillating modes” only exist for specific (eigen) frequencies.

For this case of a string stretched between two posts or walls, this set of frequencies forms a harmonic series with integer ratios between these various allowed frequencies. If we were to extend this kind of analysis to vibrating objects in two dimensions, such as cymbals or drum heads, or three dimensions, such as bells or bars that support twisting and bending modes, the resulting eigen frequencies are not necessarily in integer ratios, though such systems will generally still have eigen frequencies and corresponding eigenmodes.

Another way of looking at standing waves on a string is to view them as sums of forward and backward propagating waves. For example, the wave $A \sin(kz - \omega t)$ is a solution

¹⁸ The actual eigenvalues of Eq. (2.47) or Eq. (2.50) are technically the numbers $-k_n^2$, which obviously we can deduce from these values k_n .

of the full wave equation Eq. (2.43) (with k and ω related through Eq. (2.59)); this corresponds to a wave propagating to the “right”. The wave $A\sin(kz + \omega t)$ is also a solution, corresponding to a wave propagating to the “left”. Because this wave equation is linear, the sum or “superposition” of these two waves is also a solution. From elementary trigonometric manipulations, we can write that total wave as

$$\begin{aligned}\Psi(z, t) &= A\sin(kz - \omega t) + A\sin(kz + \omega t) \\ &= 2A\cos\omega t\sin(kz)\end{aligned}\tag{2.61}$$

So our standing wave of the form $\sin kz$ that is oscillating at frequency ω can be viewed as the sum of “right” propagating and “left” propagating waves. We can also think of these waves in terms of reflections; the right-propagating wave reflects off the right barrier to give the left-propagating wave, and the left-propagating wave reflects off the left barrier to give the right-propagating wave. Note that, in this “reflecting wave” picture, when a wave reflects off a “hard wall” barrier, the reflected wave amplitude is minus the incident wave amplitude so that the sum of the two waves always equals zero at the position of the barrier, as required for the net wave to be zero at the barrier. Equivalently, we can say that on reflection off a “hard wall”, a wave has its phase changed by 180° .¹⁹

Quiz questions

Q2.7.1 For a uniform string of a given tension stretched between one wall at position $z = 0$ and another at position $z = L$, which of the following are possible values for k for the amplitude of a standing wave solution of the form $A\sin(kz)$

- a) $k_n = \frac{n}{L}$ for $n = 0, 1, 2, \dots$
- b) $k_n = \frac{n\pi}{2L}$ for $n = 0, 1, 2, \dots$
- c) $k_n = \frac{n\pi}{L}$ for $n = 0, 1, 2, \dots$
- d) $k_n = \frac{n}{2L}$ for $n = 0, 1, 2, \dots$

Q2.7.2 For a uniform string of a given tension stretched between one wall at position $z = 0$ and another at position $z = L$, which of the following are possible forms of standing wave solutions, for appropriate non-zero values of k

- a) $A\cos(kz)$

¹⁹ If we were to reflect a wave off the open end of a string, incidentally, the phase change on reflection would be zero degrees – that is, the amplitude of the reflected wave would equal that of the incident wave. Then, instead of having a zero at the end of the string, we would have a maximum.

- b) $A \cos(kz) + B \sin(kz)$
 c) $A \sin(kz)$

Problems

2.7.1 Suppose we have a guitar string 0.5 mm in diameter, and made from steel with a density of 7.86 g/cm^3 .

- (i) What is the linear density of this guitar string (in kg/m)?

Now, we tension the string with a tension of 50 N.

- (ii) What is the wave velocity on the string (in m/s)?

Now we choose the string to be 50 cm long and attach it at both ends to rigid supports (such as the neck and the bridge of a guitar) with the tension in the string still being 50 N.

- (iii) What is the frequency (in Hz) of the lowest-frequency standing-wave resonance on this string?

2.7.2* A guitar string is 60 cm long and has a mass per unit length of $8 \times 10^{-4} \text{ kg/m}$. It is held between the two ends (i.e., at the “bridge” and the “neck”) with a tension of 80 N.

- (a) What are the frequencies of the fundamental (i.e., lowest frequency) mode and the second harmonic (i.e., the second-lowest frequency) mode?
 (b) What happens to the pitch (i.e., frequency) of a given mode if I loosen the string (i.e., reduce the tension) by adjusting the tuning peg?

2.7.3 Consider a string of length L stretched between two rigid posts. The wave velocity on the string is v . (The string is presumed to be along the “horizontal” z direction, and y corresponds to a “vertical” displacement of the string.) For each of the following functions, which represent the y displacement of the string, state whether or not the function is a possible solution of the Helmholtz equation for this string, and if so, give an expression for the angular frequency of the oscillation in time that is associated with this solution. Presume A and B are constants.

- (i) $A \sin\left(\frac{\pi z}{L}\right)$
 (ii) $A \sin\left(\frac{2.5\pi z}{L}\right)$
 (iii) $A \sin\left(\frac{9\pi z}{L}\right)$
 (iv) $A \sin\left(\frac{3\pi z}{L}\right) + B \sin\left(\frac{5\pi z}{L}\right)$

2.7.4* Consider a uniform string of length L stretched with some specific tension in the z direction between two rigid posts. Write down the spatial forms (i.e., the functions $y(z)$) for each of the three standing wave eigenfunctions with the lowest frequencies. [In other words, find the solutions of the Helmholtz equation corresponding to the lowest three eigenvalues of this problem. Note that a function that is zero everywhere will not be counted as an eigensolution.] (Your answer can be arbitrary within an overall multiplying constant; in such a problem, if some function $f(z)$ is an eigenfunction corresponding to a given eigenvalue, then so also is $Bf(z)$ where B is any constant.)

2.8 A coupled oscillator 4a

We can usefully look at one other classical problem here, which is a coupled oscillator. This will expose some more of the mathematics of modes and their properties. The physics of coupled systems also is found throughout quantum mechanics, where it underpins many basic ideas of interactions; it helps us understand chemical bonding, crystals, and properties of materials generally, and how different kinds of things like light or vibrations interact with one another or with matter.

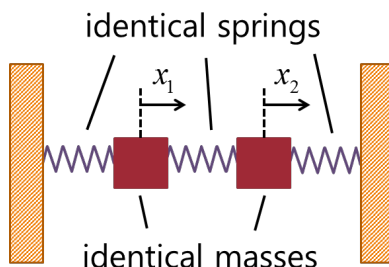


Fig. 2.6. Coupled oscillator with two identical masses and three identical springs.

A simple example of such a coupled system is two masses and three springs, as shown in Fig. 2.6. We are presuming here that the masses can only move from side to side, and that the motion is frictionless. We can view this as being two “mass-on-a-spring” simple harmonic oscillators that are additionally coupled through the spring connecting the two masses.

Equations for coupled oscillators

To analyze this problem, we need separate coordinates for the two masses. We describe the positions of the two masses relative to their equilibrium positions as x_1 and x_2 respectively. We will take both masses to be equal and of value m . For simplicity here, we will take all three springs to have the same spring constant K .

We presume that, at equilibrium, there is no net amount of stretching or compression of any of the springs²⁰. Now if we displace the left mass to the right by an amount x_1 , the force from the spring on the left on the mass on the left is $-Kx_1$, with the negative sign corresponding to pulling to the left, back towards the equilibrium position. In general, the mass on the right will also be displaced by some amount x_2 . So the spring in the middle will be stretched by an amount $x_2 - x_1$. If that spring is stretched, it is pulling the mass on the left to the right with a force $K(x_2 - x_1)$, so the net force (to the right) on the mass on the left is $-Kx_1 + K(x_2 - x_1)$. Note because of our definitions, this is the force in the positive x direction (that is, to the right) on the mass on the left²¹.

We now can apply Newton's second law to the mass on the left, giving

$$m \frac{d^2 x_1}{dt^2} = -Kx_1 + K(x_2 - x_1) = -2Kx_1 + Kx_2 \quad (2.62)$$

We can similarly analyze the force in the positive x direction on the mass on the right. Putting that into Newton's second law for the mass on the right gives

$$m \frac{d^2 x_2}{dt^2} = -K(x_2 - x_1) - Kx_2 = -2Kx_2 + Kx_1 \quad (2.63)$$

Now we see that we have two simultaneous differential equations, Eqs. (2.62) and (2.63). We want to solve these equations to show us the possible oscillating modes of this system.

Finding the eigenmodes

One of the intuitive ideas we have about a mode of oscillation is that, in such a mode, everything that is oscillating is oscillating at the same frequency. To look for such oscillating modes in this problem, we therefore propose that there are (sinusoidal) oscillating solutions at angular frequency ω , and see if we can find any solutions in such a case. For solutions of the form $\sin \omega t$ or $\cos \omega t$ for the motions of either mass, we therefore have, from (2.62) and (2.63) respectively,

$$-\omega^2 m x_1 = -2Kx_1 + Kx_2 \quad (2.64)$$

and

$$-\omega^2 m x_2 = -2Kx_2 + Kx_1 \quad (2.65)$$

Simple manipulations (swapping left and right sides and changing signs) lead to

²⁰ Actually, it makes no real difference to this problem if the springs are already compressed or stretched at their equilibrium positions here, but the algebra is slightly simpler to write down initially if we presume no stretching or compression.

²¹ Of course, this force might have a negative value, which is the same as positive force acting to move the mass to the left.

$$2Kx_1 - Kx_2 = \omega^2 mx_1 \quad (2.66)$$

and

$$2Kx_2 - Kx_1 = \omega^2 mx_2 \quad (2.67)$$

These equations can be written in a matrix form as

$$\begin{bmatrix} 2K & -K \\ -K & 2K \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \omega^2 m \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (2.68)$$

That is, the “top line” of this matrix equation is equivalent to Eq. (2.64) and the “bottom line” is equivalent to Eq. (2.65). We can choose to write, for some variable λ ,

$$\lambda = \frac{\omega^2 m}{K} \quad (2.69)$$

Using this, we can rewrite Eq. (2.68) as

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (2.70)$$

Then, we can rewrite Eq. (2.70) in the abstract mathematical notation introduced earlier as in, for example, Eq. (2.30), as

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad (2.71)$$

with

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \quad (2.72)$$

and

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (2.73)$$

We see that we have reduced this problem to an eigen problem, with eigenvalue λ . When written in the form of Eq. (2.71), it is in exactly the same general form as we encountered in dealing with the simple, one-mass, oscillator (Eq.(2.30)). Here, as before, the operator $\mathbf{A}$ is linear (in fact, matrices with constant elements only ever represent linear operators). Now $\mathbf{x}$ is a vector, with its two elements representing the positions of the two masses.

To solve the problem, we can now take the standard approach for finding the eigenvalues and eigenvectors of a matrix. Rewriting Eq.(2.70), we have

$$\begin{bmatrix} 2 - \lambda & -1 \\ -1 & 2 - \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \quad (2.74)$$

For Eq. (2.74) to hold, a standard result of linear algebra is that the determinant of the matrix must be zero, that is,

$$(2 - \lambda)(2 - \lambda) - 1 = 0 \quad (2.75)$$

So

$$\lambda^2 - 4\lambda + 3 = 0 \quad (2.76)$$

Hence

$$\lambda = 2 \left(1 \pm \frac{1}{2} \right) \quad (2.77)$$

or, equivalently

$$\omega^2 = \frac{2K}{m} \left(1 \pm \frac{1}{2} \right) \quad (2.78)$$

From Eq.(2.74), substituting the allowed values of λ , one by one, we can deduce the form of the associated eigenvectors, which are, respectively,

$$\lambda = 1: \quad x_1 = x_2, \text{ i.e., } \mathbf{x} \propto \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (2.79)$$

which corresponds to the two masses oscillating in the same direction, with equal amplitude, with frequency $\omega = \sqrt{K/m}$, and

$$\lambda = 3: \quad x_1 = -x_2, \text{ i.e., } \mathbf{x} \propto \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad (2.80)$$

which corresponds to the two masses oscillating in opposite directions, with equal (but opposite) amplitudes, with frequency $\omega = \sqrt{3K/m}$. Perhaps not surprisingly, then, now that we have two masses, we have discovered two different modes, a symmetric, lower frequency one, with both masses going in the same direction, and an antisymmetric, higher frequency one with them going in opposite directions.

Quiz Question

Q2.8.1 For the two masses and three springs as above:

- The frequency of oscillation of the lower frequency coupled oscillator mode is the same as the frequency of oscillation we would have for a single one of the masses hanging from a single one of the springs. True or False?
- if we replaced the spring in the middle with a much stiffer one, the frequency of oscillation of the lower frequency coupled oscillator mode would stay the same. True or False? (Hint: think about how much this spring is stretched in lower frequency mode in the “equal springs” case analyzed in the text.)
- if we doubled the size of both masses in the two-mass/three-spring system as in the lecture, the difference between the two mode frequencies (i.e., in Hz) would remain the same. True or False?

Problems

- 2.8.1 Consider a coupled oscillator similar to the one shown in as shown in Fig. 2.6. The two masses are identical, each with mass m , and the slide horizontally without friction. The left and right springs have the same spring constant k_1 , but the middle spring now has a possibly different spring constant k_2 . We can take the three springs all to have no net extension or compression at the equilibrium position. The displacements of the masses in the x direction relative to their equilibrium positions are x_1 and x_2 , respectively, as shown in Fig. 2.6.

Derive an expression or expressions for the possible resonant (angular) frequencies (i.e., (angular) eigen frequencies) of this coupled oscillator system.

- 2.8.2 For the coupled oscillator in the preceding problem, calculate the eigenvectors of the matrix corresponding to this system, and describe or draw how the masses are moving in time relative to one another in each corresponding mode of oscillation.

2.9 Mathematical properties of modes

The most obvious property of the modes we have been discussing is that they describe oscillations at specific frequencies. In linear systems, however, the mathematics behind modes turns out to be quite deep and quite generally useful well beyond those simple situations. Here we look further at some of that mathematics and its use in describing the physics of linear systems. We use some of the results from the coupled oscillator to illustrate these mathematical properties.

Inner products and orthogonality

We remember from geometrical vectors that we can take the “dot” product between two vectors. We could, for example, have two vectors in the x - y plane, such as $\mathbf{a} = 2\mathbf{i} + 3\mathbf{j}$ and $\mathbf{b} = -5\mathbf{i} + 4\mathbf{j}$ where $\mathbf{i}$ and $\mathbf{j}$ are the unit vectors in the x and y directions, respectively. The vector dot product between these vectors is the sum of the products of the components, that is

$$\mathbf{a} \cdot \mathbf{b} = 2 \times (-5) + 3 \times 4 = 2 \quad (2.81)$$

We also remember that it is possible to multiply two vectors written out in matrix-vector notation provided we take some care with matching the “dimensions” of the vectors being multiplied. Specifically, we can multiply a column vector on the “right” and a row vector on the “left” if the column and the row are the same “length” – that is, if they have the same number of elements. We could choose to write out the geometrical vector dot product in this format, making $\mathbf{a}$ a row vector and $\mathbf{b}$ a column vector, so we would have

$$\mathbf{a} \cdot \mathbf{b} = \begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} -5 \\ 4 \end{bmatrix} = 2 \times (-5) + 3 \times 4 = 2 \quad (2.82)$$

If we choose two geometrical vectors that are at right angles to one another, such as $\mathbf{c} = 2\mathbf{i} + 3\mathbf{j}$ and $\mathbf{d} = -3\mathbf{i} + 2\mathbf{j}$, then their dot product is zero, as we can see.

$$\mathbf{a} \cdot \mathbf{b} = 2 \times (-5) + 3 \times 4 = 2 \quad (2.83)$$

Of course, these ideas of products of vectors are valid even when the vectors are not ones that are also vectors in ordinary geometrical space. In that case, we can call products of row and column vectors as in Eq. (2.82) “inner products” as a generalization of the idea of the geometrical vector dot product. If the inner product of two non-zero vectors is zero, we say the vectors are “orthogonal” as a generalization of the idea of the vectors being at right angles.

Now, if we look at the two eigenvectors we found from analyzing the coupled oscillator problem above – that is, the eigenvectors in Eqs. (2.79) and (2.80) – and we take the inner product between them, we find that inner product is zero; they are mathematically orthogonal. Explicitly,

$$\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 0 \quad (2.84)$$

If we had drawn these two vectors on a piece of paper as geometrical vectors, as in Fig. 2.7, we would find they are at right angles to one another.

This orthogonality property of these modes is no accident; in fact, for a very broad class of the linear physical problems of interest to us, the eigenfunctions or (eigen) modes are orthogonal in this sense. In particular, this is true in a large and important range of cases in the physical descriptions of quantum mechanics, and we will discuss this below.

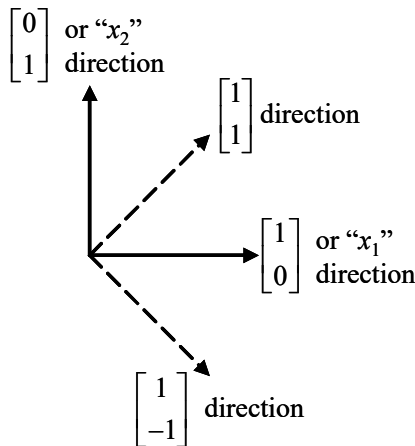


Fig. 2.7. Plot of the eigenvectors of the coupled oscillator.

Functions as vectors

Note now that we are starting to think of a function as a vector in some kind of mathematical space. Indeed, we could represent any particular pair of positions of the two masses in the coupled oscillator as a vector on the diagram in Fig. 2.7. The function

here is a very simple one, and can be written simply as a list of two numbers, corresponding to the position of mass 1 and the position of mass 2.

Any list of numbers can be written as a vector. We can represent any function as simply a list of numbers, representing the results of the function when mapping from a known list of arguments²². For functions of continuous variables that list would be infinitely long, but the concept remains usable with some care even in that case. In general, though, we can if we want represent any function as a vector. We will give more examples of this idea below.

Completeness and basis sets

Another very important point about these eigenfunctions or eigenmodes is that they can be used to describe any particular position of the two masses. This rather important property that sets of modes can have is called “completeness”. A particular pair of positions of the two masses, namely $x_1 = f$ and $x_2 = g$ can rather obviously be represented as

$$\begin{bmatrix} f \\ g \end{bmatrix} = f \begin{bmatrix} 1 \\ 0 \end{bmatrix} + g \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (2.85)$$

with f and g just being the “components” along the “ x_1 ” and “ x_2 ” directions respectively in Fig. 2.7. We could, however, equally well write this same pair of positions as the components along the directions corresponding to the two eigenvectors; that is, we could write

$$\begin{bmatrix} f \\ g \end{bmatrix} = \frac{(f+g)}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{(f-g)}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad (2.86)$$

So, we could say that the set of eigenvectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ is complete for representing any possible pair of positions of the two masses.

If we have such a set of vectors that can be used to make up any vector in the “space” of interest (here a 2-dimensional “plane”), we can call that set a “basis set” of vectors. If the vectors in the basis set are all orthogonal to one another, then we can call the set an “orthogonal basis set” (or often just an “orthogonal basis”).

²² The list of values of the argument of the function is usually implicit, though ideally we should be more explicit about it. In the present case, we could say we are mapping from an “index” that has two values – “1” for the mass on the left and “2” for the mass on the right – to the corresponding positions of the masses, x_1 being the position of mass “1” and being the first element of the vector, and x_2 being the position of mass “2” and being the second element in the vector. It is a weakness of the vector notation for specifying functions that the corresponding values of the argument are not written out explicitly, and we simply have to deal with that by some other notational means whenever there is any confusion.

It is obvious if we look at Fig. 2.7 that we could equally well represent any point in the plane in terms of its coordinates along the $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ directions, or in terms of its coordinates along the $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ directions. Either one of these “orthogonal basis sets” is equally usable. Another way of stating this same property is that the modes form a complete set for “spanning” the space²³ that describes the positions of the masses.

Hermitian adjoints

If we look at the operator A in this problem, as given by Eq. (2.72), we can see that the matrix is symmetric if we reflect it along its leading diagonal (the diagonal with the elements of value 2). In the general mathematics of modes, it is usual to allow for the possibility of complex quantities, and then, in an extension of the idea of a symmetric matrix, we can define a Hermitian matrix.

First, we need formally to define the idea of a Hermitian adjoint of a matrix. The operation of taking a Hermitian adjoint involves reflecting a matrix along its leading diagonal (i.e., taking the transpose of the matrix), and taking the complex conjugate of the elements when we do so. The Hermitian adjoint is often denoted with the “dagger” symbol “ $\dagger$ ”. For example, for a 2×2 matrix, we have

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{\dagger} \equiv \begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} \quad (2.87)$$

We can also have Hermitian adjoints of matrices that are not square, with the most common example being a vector. In the case of a 2-element vector, for example, we have

$$\begin{bmatrix} a \\ b \end{bmatrix}^{\dagger} \equiv \begin{bmatrix} a^* & b^* \end{bmatrix} \quad (2.88)$$

Again, we can think of this operation as reflecting about a “45°” diagonal line (from top left to bottom right) and taking the complex conjugate of the elements.

²³ Incidentally, the “space” we are speaking about here is not usually an actual geometrical space, but rather a function space, i.e., one in which we can represent entire functions. (Here our example functions are rather simple, since they can be described entirely by specifying only two numbers, and so we can represent them in a two-dimensional space.) Spaces of this class that are of most interest to us are known as Hilbert spaces. They are mathematically like ordinary geometrical spaces in many ways, but are different, first, in that they may have any number of dimensions, including an infinite number, because that many values may be required to represent a function, and, second, that the coordinates in such a space may be complex numbers rather than only real numbers.

We can think of the Hermitian adjoint as being like the idea of a complex conjugate but now generalized to matrices or operators. A simple scalar number can be thought of as a 1×1 matrix, and the Hermitian adjoint of that “matrix” is simply the complex conjugate of the number.

Given that we now have the possibility of complex numbers in our mathematics, we need to make one other minor change. We need to extend the “inner product” to being one in which we use the Hermitian adjoint vector as the one on the left²⁴, i.e., the inner

product of $\begin{bmatrix} a \\ b \end{bmatrix}$ and $\begin{bmatrix} c \\ d \end{bmatrix}$ is given by

$$\begin{bmatrix} a^* & b^* \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = a^*c + b^*d \quad (2.89)$$

Dirac notation for complex vectors

Now that we are dealing with complex vectors, we sometimes need a good notation for working with them. We will not make much use of this here, but so that we have a notation when we do need it, we can introduce the one used most in physics, which is called Dirac²⁵ notation, or “bra-ket” notation. For our purposes, a column vector of numbers which we can write as $\mathbf{x}$ in one of our abstract notations is written $|\mathbf{x}\rangle$ in Dirac notation and is then called a “ket” vector. This is the notation on the far right here

$$\mathbf{x} \equiv \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} \equiv |\mathbf{x}\rangle \quad (2.90)$$

The Hermitian adjoint of the vector $\mathbf{x}$, which could be written as $\mathbf{x}^\dagger$ in one of our notations, is, in the Dirac notation, the “bra” vector $\langle \mathbf{x}|$, which is the notation on the far right here

²⁴ The reader might be asking why do we need to use this Hermitian adjoint, with its complex conjugates, when taking this inner product? The reason is that we need to have a simple measure of “length” (for a vector) that gives a real number, even if the elements of the vector are not real. In the language of the mathematics of these Hilbert spaces, we need a “norm” that is real. It is very useful (and in fact a requirement for a Hilbert space) if that norm is constructed from the inner product of the vector with its own adjoint (in fact, the norm is defined as the square root of that inner product, just as it would be for vectors in an ordinary geometrical space). Taking the complex conjugate in constructing the adjoint guarantees the result for the “norm” or “length” is a real number, just as taking the product of a complex number with its complex conjugate always results in a real number.

²⁵ Paul Dirac (1902 – 1984) was a British theoretical physicist and one of the most outstanding theoretical physicists of the 20th century. He made several major contributions to quantum mechanics in particular, including a relativistic version of quantum mechanics that effectively explained spin and introduced antiparticles, and the quantum mechanics of light and its interaction with matter. Incidentally, his first university degree was in electrical engineering.

$$\mathbf{x}^\dagger \equiv \begin{bmatrix} x_1^* & x_2^* & \cdots \end{bmatrix} \equiv \langle \mathbf{x} | \quad (2.91)$$

and

$$| \mathbf{x} \rangle^\dagger = \langle \mathbf{x} | \quad \text{and} \quad \langle \mathbf{x} |^\dagger = | \mathbf{x} \rangle \quad (2.92)$$

which is obvious if we think of these as column or row vectors as appropriate. The real reason for having and using the Dirac notation is that it gives a clear distinction between what are essentially column vectors and what are essentially row vectors. Writing a vector just as $\mathbf{x}$ does not make this explicitly clear.

With the Dirac notation, we now have a definite and clear way of writing the inner product between two (possibly complex) vectors $| \mathbf{x} \rangle$ and $| \mathbf{y} \rangle$, which we can write as

$$\langle \mathbf{x} | \mathbf{y} \rangle \equiv \begin{bmatrix} x_1^* & x_2^* & \cdots \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix} \quad (2.93)$$

Note that, though we could write this inner product as $\langle \mathbf{x} || \mathbf{y} \rangle$, which would be correct and would cause no problems, conventionally we drop the second vertical line and write $\langle \mathbf{x} | \mathbf{y} \rangle$. Incidentally, we can note that, quite generally for such inner products

$$\langle \mathbf{x} | \mathbf{y} \rangle = \langle \mathbf{y} | \mathbf{x} \rangle^* \quad (2.94)$$

which is easy to check based on viewing the vectors as column and row vectors as in Eq. (2.93). Note that the result of such an inner product here is just generally a complex number, and the superscript “*” just refers to the complex conjugation of that number as usual. Eq. (2.94) also helps us write down a key property we require for such inner products with complex functions or vectors, which is that “flipping” the order of the vector in the inner product results in having to take the complex conjugate.

This Dirac notation version of writing the inner product is particularly clear and intuitive²⁶. The inner product here even has a kind of “inner” look to it with the angled brackets “enclosing” the expression.

We can also continue to use Dirac notation when we include operators in our expressions. So we can rewrite Eq. (2.71) ($\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$), for example, as

²⁶ There is another very unfortunate notation often used in mathematical texts for the inner product, which is to write for the inner product $\langle \mathbf{x} | \mathbf{y} \rangle$ instead the notation $(\mathbf{y}, \mathbf{x})$. Note that this is the “wrong” way round. Now the associative law as in matrix-vector products no longer is visually obvious. In Dirac notation, we can write, for example, $\langle \mathbf{x} | \mathbf{A} | \mathbf{y} \rangle \equiv (\langle \mathbf{x} | \mathbf{A}) | \mathbf{y} \rangle \equiv \langle \mathbf{x} | (\mathbf{A} | \mathbf{y} \rangle)$ which is just using the normal associated property of matrix-vector multiplications that allows us to group them in any convenient way as long as we do not change the order. Trying to write the same equivalence in the mathematical “ $(\mathbf{y}, \mathbf{x})$ ” notation is extremely confusing and far from obvious ($(\mathbf{y}, \mathbf{A}\mathbf{x}) \equiv (\mathbf{A}\mathbf{y}, \mathbf{x})$).

$$A|\mathbf{x}\rangle = \lambda|\mathbf{x}\rangle \quad (2.95)$$

Dirac notation is widely used in more advanced treatments of quantum mechanics.

Hermitian matrices and operators

A Hermitian matrix is one that is equal to its own Hermitian adjoint. In abstract operator notation, a Hermitian operator is one for which

$$A = A^\dagger \quad (2.96)$$

For example, the matrices

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & i \\ -i & -2 \end{bmatrix}, \text{ and } \begin{bmatrix} 0 & 1-i & -3 \\ 1+i & -1 & \exp(i/\sqrt{2}) \\ -3 & \exp(-i/\sqrt{2}) & 1 \end{bmatrix}$$

are all Hermitian, but the matrices

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} i & 1 \\ 1 & -i \end{bmatrix}, \text{ and } \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$$

are not.

In general, essentially²⁷ any linear Hermitian operator can be represented by a Hermitian matrix. Hermitian matrices have three very useful properties:

- (i) their eigenvectors are always orthogonal to each other;
- (ii) the eigenvalues are always real; and
- (iii) the eigenvectors form a complete (and orthogonal) set for the space in question.

Because all the elements of the matrix A as given by (2.72) happen to be real, then this particular symmetric matrix is also Hermitian.

Hermitian operators and physical problems

In practice, for a large number of physical problems associated with linear systems, especially those that correspond to frictionless or loss-less systems, the key linear operators we deal with are Hermitian²⁸, and can be represented to any degree of

²⁷ Technically, any “compact” Hermitian operator can certainly be represented by a matrix to any desired degree of accuracy. In practice, this typically includes all bounded operators, that is, ones that when they operate on a bounded function lead to a bounded function as a result.

²⁸ As the reader may have noticed, the operator d^2/dt^2 is Hermitian, and the rather trivial operation of multiplying a function by a real constant is also Hermitian. Somewhat surprisingly, the operator d/dt is not Hermitian, though the operator id/dt is. Simple frictional systems can have terms in the differential equation that involve the operator d/dt because frictional damping

accuracy we wish by a sufficiently large Hermitian matrix²⁹. In particular, a large number of the operators we deal with in quantum mechanics are of this type³⁰.

By extension from the matrix case, the eigenfunctions of Hermitian operators that can be represented by matrices are always orthogonal, with real eigenvalues, and form a complete set of basis functions for the relevant space.

Hence, not only are eigenfunctions and eigenmodes interesting for the physical behaviors they describe, such as the distinct ways that something can oscillate, but also, for Hermitian operators representing physical systems, these eigenfunctions or eigenmodes can have rather remarkable and useful properties, specifically, orthogonality and completeness; as a result, we can use them mathematically to describe all sorts of behaviors of physical systems, in ways that often turn out to be very convenient.

Standing waves and Fourier series

As we mentioned above, the sets of modes that we get by solving a particular physical problem can also turn out to be complete sets for describing functions in the relevant function space. If we return to the problem of the standing wave modes on a string, our function space is one containing the one-dimensional functions of position defined between 0 and L (and having the value zero at positions 0 and L). If our sets of modes are complete, we should be able to describe any functions between positions 0 and L (at least if the functions are zero at 0 and L) in terms of combinations of these modes.

The reader may be aware of Fourier series. There are a few different forms of Fourier series, but essentially Fourier series can be used to represent arbitrary functions as sums of sine waves. To represent an arbitrary function $f(z)$ between 0 and L (at least one that is zero at 0 and L) we could use a Fourier series of the form

$$f(z) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi z}{L}\right) \quad (2.97)$$

force is often proportional to velocity, at least in simple systems. Physical systems that have such frictional damping are therefore not represented by Hermitian operators.

²⁹ Operators that can be represented to any desired degree of accuracy by a sufficiently large matrix are essentially what are called “compact” operators, though the actual definition of compact operators is somewhat technical mathematically, and requires ideas from the subject known as functional analysis. Functional analysis also gives us the kinds of results that allow us to generalize the ideas here to infinite dimensional spaces and from finite matrices to differential equations. This subject can be rather forbidding, unfortunately, though see D. A. B. Miller, “An introduction to functional analysis for science and engineering,” <http://arxiv.org/abs/1904.02539> for a relatively readable introduction.

³⁰ In particular, all the operators associated with what are called measureable quantities in quantum mechanics are Hermitian.

where the a_n are appropriate numbers (coefficients). The set of coefficients a_n is just as good a way of describing the function as is the list of values of the function for all values of z of interest.

Our modes here for standing waves on a string are actually identical to the sine waves used in the Fourier series. Hence, since we already know that Fourier series can represent arbitrary functions in this way, so also can our set of modes. In other words, from our knowledge of Fourier series, we already know that this particular set of functions forms a complete set, a property we have claimed that sets of modes possess. Alternatively, we could state that the Fourier basis functions (the sine waves) form a complete set because they happen to be the eigenfunctions of a Hermitian operator.

Generality of basis sets of functions

Note, incidentally, that such basis sets of functions can be used to represent any function, regardless of whether the actual physical situation corresponds to the one used in deriving the set. For example, we could have a string whose density varied along its length. The modes of such a string would not be the simple sine waves we have here. Nevertheless, at any given time, we could if we wanted still represent the shape of the string as a sum of our sine wave modes. In such a case, we are using the mathematical properties of the set of eigenfunctions of an operator even though that operator is not necessarily the one that corresponds to the current physical system's behavior.

There is actually an infinite number of possible basis sets of functions to describe functions in any given space. This mathematical fact can be very useful; in any given problem, there is often a very convenient basis to choose that makes the problem simple to solve. This notion of multiple different possible sets of basis functions is also basic to the mathematical foundation of quantum mechanics.

Modes as the eigenfunctions of operators

Finally, we can write a definition that will encompass the concept of modes for our purposes. As the reader has been warned, this is rather mathematical, but we have now introduced enough terminology that we can make the statement.

The broadest possible statement we can make is

$$\text{A mode is an eigenfunction of an operator that describes a physical system.} \quad (2.98)$$

The system need not necessarily be linear, and even if linear, it need not be represented by Hermitian operator.³¹

³¹ There are some operators that are not Hermitian that lead to real eigenvalues (in the problem of two unequal masses with three springs, the resulting matrix is not Hermitian, though the eigenvalues are real)

The most useful sets of modes from a mathematical point of view are those that correspond to linear, Hermitian operators.³²

We can also state³³ that

For linear physical systems that can be described by a Hermitian operator that can be approximated to any degree of accuracy by a finite matrix,

the modes are the eigenfunctions of the operator,

they are orthogonal,

they are complete, and

they have real eigenvalues.

The modes we have discussed so far have been what we could call “oscillating modes”. Such modes each have a specific frequency associated with them, and the eigenvalue associated with a given mode is related to that frequency. For such oscillating modes we can say

In an oscillating mode, everything that is oscillating is oscillating at the same frequency.

This is often in practice a good way to look at oscillating modes.

Quiz questions

Q2.9.1 For each of the following pairs of vectors, state whether or not the vectors are orthogonal.

- a) $x_1 = [1, -1]$ and $x_2 = [1, 0]$
- b) $x_1 = [2, -3]$ and $x_2 = [6, 4]$
- c) $x_1 = [1, -1, 2]$ and $x_2 = [1, -1, -1]$
- d) $x_1 = [1, 0, 0]$ and $x_2 = [0, -1, 0]$

Q2.9.2 For each of the following pairs of vectors, state whether the pair forms a complete orthogonal basis for describing two-dimensional vectors in a plane (e.g., as in the diagram in the text)

- a) $x_1 = [-3, 2]$ and $x_2 = [2, 3]$

³² A broader mathematical definition of modes of linear systems would be to say that they are the solutions of the generalized eigenvalue problem. Such a problem can be written in the form $A\mathbf{x} = \lambda B\mathbf{x}$ where A is a linear Hermitian operator and B is a linear operator also.

³³ This definition is slightly imprecise in that not necessarily all Hermitian operators can be represented to any degree of accuracy by finite matrices, though all “compact” ones can. In practice, however, we can typically reduce linear problems to ones that we can solve with matrices, and in that sense we can regard this definition as being broad enough for a discussion of such Hermitian problems.

b) $x_1 = [6, 18]$ and $x_2 = [-2, -6]$

Q2.9.3 For each of the following pairs of vectors, state whether or not the inner product is zero.

a) $|\mathbf{x}_1\rangle = \begin{bmatrix} 1 \\ -i \end{bmatrix}$, $|\mathbf{x}_2\rangle = \begin{bmatrix} -1 \\ -i \end{bmatrix}$

b) $|\mathbf{x}_1\rangle = \begin{bmatrix} 1+i \\ -1 \end{bmatrix}$, $|\mathbf{x}_2\rangle = \begin{bmatrix} 1+i \\ 2 \end{bmatrix}$

Q2.9.4 For each of the following matrices, state whether or not the matrix is Hermitian.

a) $H = \begin{bmatrix} i & -1 \\ -i & 3 \end{bmatrix}$

b) $H = \begin{bmatrix} 1 & i & 0 \\ i & -8 & -i \\ 0 & -i & 0 \end{bmatrix}$

c) $H = \begin{bmatrix} 8 & 1-i & 0 \\ 1+i & -8 & -e^{i\alpha} \\ 0 & -e^{-i\alpha} & 8 \end{bmatrix}$

Problems

2.9.1 Consider the following matrices and state whether they are Hermitian.

(i) $\begin{bmatrix} 1+i & 1 \\ 1 & 1-i \end{bmatrix}$

(ii) $\begin{bmatrix} 7.5 & 3-i \\ i-3 & 20 \end{bmatrix}$

(iii) $\begin{bmatrix} 14 & 1+7i \\ 1-7i & 5 \end{bmatrix}$

(iv) $\begin{bmatrix} |z| & \exp(-i\theta) & i \\ \exp(i\theta) & -|z| & 2-3i \\ -i & 2+3i & 0 \end{bmatrix}$ where z is a complex number and θ is a real number.

2.9.2 Consider the following pairs of vectors and state whether they are orthogonal. (Use the form of the inner product appropriate for complex-valued vectors.)

(i) $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} -4 \\ 2 \end{bmatrix}$

(ii) $\begin{bmatrix} 1 \\ i \end{bmatrix}$ and $\begin{bmatrix} i \\ 1 \end{bmatrix}$

(iii) $\begin{bmatrix} 1+i \\ 3 \end{bmatrix}$ and $\begin{bmatrix} -3 \\ 1+i \end{bmatrix}$

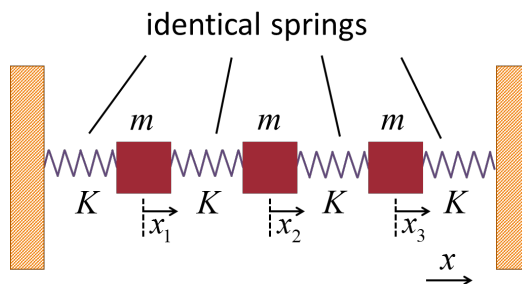
2.9.3 For the matrices listed below, identify which are Hermitian. For any Hermitian matrices, verify that their eigenvalues are real and their eigenvectors are orthogonal to each other.

(a) $\begin{bmatrix} 2 & 2+i & 3 \\ 2+i & 1 & i \\ 3 & i & 2 \end{bmatrix}$

(b) $\begin{bmatrix} 2i & 1 \\ 1 & -2i \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 2i \\ -2i & -1 \end{bmatrix}$

2.9.4 Consider three masses and four springs as in the figure. All the masses are identical, with mass m . All the springs are identical, with spring constant K . The masses are presumed to be free to move in the x direction, with no friction, but they cannot move in any other direction. The positions of the masses with respect to their equilibrium positions are x_1 , x_2 , and x_3 respectively. At the equilibrium positions (which correspond, respectively, to $x_1 = 0$, $x_2 = 0$, and $x_3 = 0$), we can take all the springs to have no extension or compression.



For any given set of positions, we can, if we like, write these in the form of a mathematical vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

a) By considering the total force on each mass, write down the three differential equations, one for each mass, that follow from Newton's

second law and the extension of the springs, in a similar fashion to the two-mass problem in the text.

- b) Now presume we are looking for solutions that correspond to only one (angular) frequency ω of oscillation for everything that is oscillating. Write down the resulting three equations, one for each mass, in this case.
- c) Now write these three equations from (b) in the form of a matrix eigen equation – that is, in the form $A\mathbf{x} = \lambda\mathbf{x}$. Do this in such a way that the elements of the matrix A are dimensionless (that is, they are just numbers) and λ is a positive (and also dimensionless) quantity.
- d) For each of the following vectors, state whether it is an eigenvector of A , and, if so, what is the associated eigenvalue λ and a corresponding expression for the eigen (angular) frequency ω . [Note: do *not* formally solve for the eigenvectors of the matrix, either analytically or using any computer program here – justify your answers by directly showing the vectors are solutions of the matrix eigen equation for some value of λ .] [Hint: At least two of these are eigenvectors.]

$$(i) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}; \quad (ii) \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}; \quad (iii) \begin{bmatrix} -1/2 \\ 1/\sqrt{2} \\ -1/2 \end{bmatrix}; \quad (iv) \begin{bmatrix} 1 \\ -\sqrt{2} \\ 1 \end{bmatrix}; \quad (v) \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \end{bmatrix}$$

e)

- (i) Find another eigenvector that is orthogonal to all eigenvectors in (d) above. Do this *without* solving for the eigenvectors of the matrix by any mathematical technique or program. You should be able to make an intelligent guess here and justify that your guess is indeed such an eigenvector. [Hint: In this eigenmode, one of the masses does not move at all.]
- (ii) Give the eigenvalue λ and the corresponding expression for the eigen (angular) frequency ω .
- (iii) Explicitly show the orthogonality of this vector and the other eigenvectors found in (d).

2.9.5* Suppose we have two masses and three springs, in an arrangement similar to Fig. 2.6 with the masses allowed to move without friction in one direction. As in that figure, all the spring constants are the same, with some value K , but now the mass, m_L , on the left will be different from that on the right, m_R . We can usefully define the ratio of these masses as

$$\rho = \frac{m_L}{m_R}$$

We will construct the eigen problem that would allow us to solve for the (angular) eigenfrequencies ω and the associated eigenvectors that give the relative amplitudes of the motions x_1 of mass m_L and x_2 of mass m_R .

- (i) Show that the associated eigen problem can be written as

$$M \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

where the matrix M and the quantity λ are given by

$$M = \begin{bmatrix} 2 & -1 \\ -\rho & 2\rho \end{bmatrix} \text{ and } \lambda = \frac{m_L \omega^2}{K}$$

- (ii) State whether this matrix is Hermitian, briefly justifying your answer.
- (iii) Find an expression for the eigenvalues λ of this system in terms of ρ .
- (iv) For $\rho = 2$ (so physically, $m_L = 2m_R$), write out the resulting eigenvalues λ . (The resulting expression should only involve numbers, and you may leave square roots in your expression – you need not evaluate the square roots of any numbers.)
- (v) For each of the eigenvalues, solve for the eigenvectors. Write all the eigenvectors in the form $\begin{bmatrix} 1 \\ a \end{bmatrix}$ where a is some number or numerical expression (which can include square roots that you need not evaluate, and which will generally be different for the different eigenvectors). [Note: you may find the identity $b^2 - c^2 \equiv (b - c)(b + c)$ useful if you are simplifying expressions.]
- (vi) State whether the resulting eigenvectors are orthogonal, justifying your answer briefly.

2.10 Conclusions

In this chapter, we have introduced many ideas that are well known from classical physics. The idea of eigenmodes is in particular quite widely used there, and it is also relatively well known that eigenmodes of Hermitian operators have these almost magical mathematical properties of orthogonality and completeness.

In quantum mechanics, these ideas of linear Hermitian operators, and the facts that their eigenvalues are real and that their eigenfunctions are orthogonal and form complete basis sets, are quite central to the way that we think about the whole subject and how we write it down. Though there is much that is very different about quantum mechanics compared to classical mechanics, these mathematical ideas of modes and eigenfunctions carry over very well. Understanding them in the classical context prepares us to start thinking about quantum mechanics.

Chapter 3

The quantum view of the world

3.1 The beginning of quantum mechanics

In the period leading up to 1900, the list of phenomena for which we had no good “classical” explanation was growing steadily.

By 1870, we had the first drafts of the periodic table of the elements, but we had no explanation for them or their chemical or physical properties, and indeed no underlying theory of why materials had particular properties.

The ideas of what we now call statistical mechanics were emerging in the 1870s, starting with James Clerk Maxwell’s and Ludwig Boltzmann’s work on understanding the ideas of thermal distributions. These could give some explanations of properties of gasses and, increasingly, other phenomena, such as specific heat. From Boltzmann’s work in particular, we started to have a meaningful interpretation of the difficult concept of entropy through the statistics of the occupation of possible states. One underlying presumption in such theories becomes the idea that there may be discrete states for a system; the counting of these leads to the statistical ideas of entropy. But in this early work in the 1870’s, there was no physical basis for arguing that there should be such discrete states. Indeed the ideas of atoms and molecules themselves as discrete entities were still not universally accepted.

Work on the spectroscopy of atoms such as hydrogen led to the observation of remarkable but puzzling structure in the colors of light emitted by excited hydrogen. Johann Balmer in 1885 noticed that a simple expression could predict the wavelengths of the spectral lines of hydrogen, as measured by Anders Ångström, for example, in what is now called the Balmer series of lines as shown in Fig. 3.1.

Another puzzling observation from this period was what we now call the photoelectric effect by Heinrich Hertz in 1887. When ultraviolet light shines on a metal, it emits electrons more easily (as we now understand the effect). Though classical physics might

reasonably explain why absorbing energy in the metal might lead to more emitted current, it does not explain why the wavelength of the light matters.

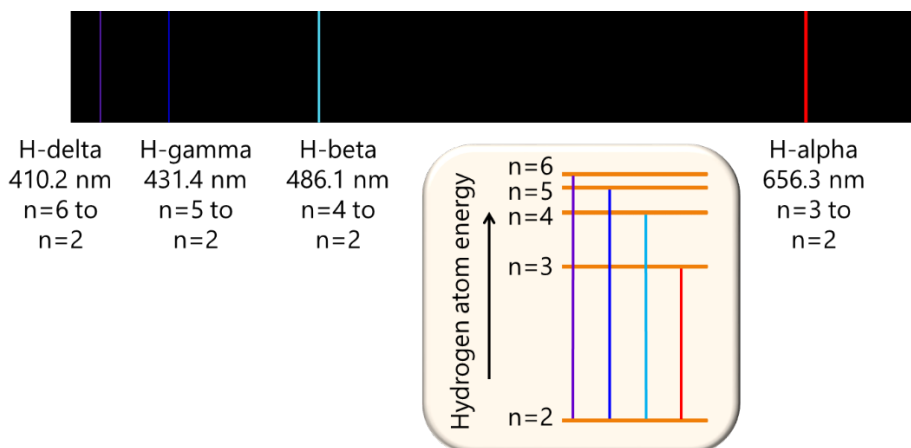


Fig. 3.1. Hydrogen atom spectral lines (top) in the visible spectrum (the Balmer series), together with their conventional names and wavelengths, and the modern identification as transitions between energy levels in the hydrogen atom.

But what in the end led to the first key conceptual breakthrough in setting up quantum mechanics was an everyday problem – lighting. Prior to the 19th century, the only practical man-made source of light was from flames. It was, of course, well known that hot objects emit light even if they themselves were not burning. Blacksmiths could effectively gauge temperature by the color of the heated metal with which they were working, for example.



Fig. 3.2. A carbon filament light bulb at the Fire Station in Livermore, California, where it has been running essentially continuously since 1901.

There had been experiments through much of the 19th century on incandescent light emitters (that is, ones that emit light because just they are hot). Electric lighting had become commercially feasible in about 1880 with carbon filament light bulbs, such as

those marketed by Thomas Edison and others (Fig. 3.2). But there was no good model for absorption and emission of light, and so it was not clear how to make lamps that were better emitters¹.

The idea that a body that was a very good absorber of light (a “black body”) would also be a very good emitter of such thermal radiation was well understood by that time, through the work of Gustav Kirchhoff and others around 1860. Kirchhoff’s Law of Thermal Radiation tells us that a body that perfectly absorbs light will be the best possible thermal emitter of light. But there was no good physical model for the spectrum of light emitted by such a black body.

In 1887, the German government, stimulated in part by Werner von Siemens, founded the Physikalisch-Technische Reichsanstalt (PTR) – the Imperial Institute of Physics and Technology – near Berlin. One priority for the PTR was to understand light emission from hot bodies for better light bulbs. By the late 1890’s, very precise measurements were being made there of the emission spectrum of a black body. Though there were some empirical formulas for portions of the spectrum, there was none that was perfect, and none of those formulas had a good physical model underlying them.

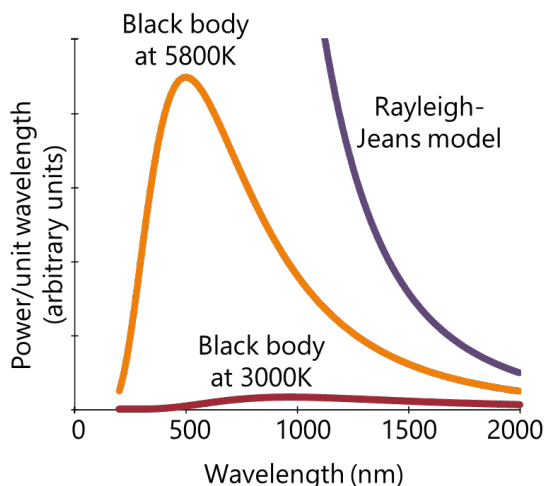


Fig. 3.3. The emission spectrum of a black body at 5800 K (approximately the temperature of the sun) and at 3000 K (the temperature of a very hot light bulb filament), and the classical Rayleigh-Jeans model.

The spectrum of emission by a black body is shown in Fig. 3.3 for two different temperatures, together with a comparison to a classical model derived by Lord Rayleigh 1900 and extended with James Jeans in 1905. First, we see that the classical model does not work at short wavelengths. This classical model predicts an ever-increasing amount of emitted light at shorter wavelengths. This failure of the classical model is therefore

¹ At that time, improving light bulbs would largely have been a practical issue of being able to achieve very high temperatures in a filament while still allowing a useably long lifetime for the light bulb; it was likely quite clear empirically that hotter bodies emitted more light that was also “whiter”. But, still, there was no underlying theory of light emission to guide progress.

sometimes known as the “ultraviolet catastrophe”. Second, we see from Fig. 3.3 and Fig. 3.4 that, since the black body is the best possible thermal emitter of light, there is a practical problem of making efficient thermal light emitters for the visible part of the spectrum (about 400 nm – 700 nm wavelength). To get the peak of the emitted spectrum to coincide with the visible range, we would need a very hot material, essentially as hot as the sun, and if we cannot make a material as hot as the sun, not only does the wavelength of peak emission move out of the visible range, the magnitude of the emission also decreases overall quite rapidly with temperature.

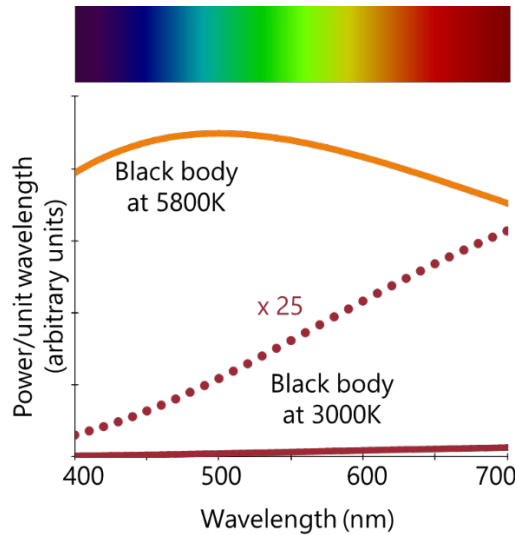


Fig. 3.4. Expanded view of the black-body spectrum at 5800 K and at 3000 K in the visible range of the spectrum. The dotted line is the 3000 K spectrum magnified by 25 times. An approximate color map of the visible spectrum is shown at the top of the figure.

Max Planck, then a professor in Berlin, in 1900 proposed a formula and then a derivation for the black-body spectrum using one “ad hoc” assumption: Suppose that light is emitted in quanta of energy

$$E = hf \quad (3.1)$$

where f is the frequency (in Hz or cycles per second) of the electromagnetic radiation. The quantity h is what has become known as Planck’s constant

$$h \approx 6.62606957 \times 10^{-34} \text{ Js} \quad (3.2)$$

Commonly, this constant is also used in the form known as “h bar”, which is Planck’s constant divided by 2π

$$\hbar \approx 1.054571726 \times 10^{-34} \text{ Js} \quad (3.3)$$

in which case we can write the energy of these quanta as

$$E = \hbar\omega \quad (= hf) \quad (3.4)$$

where ω is the corresponding angular frequency (i.e., $\omega = 2\pi f$). (It is also common to use the notation $E = h\nu$, with the Greek letter ν (“nu”) instead of f for ordinary frequency.) When Planck’s constant appears in some physical formula, we can be fairly sure we are working in the quantum world, one quite different from the classical ideas that preceded it.

Planck’s formula agrees very well with observed spectra. Note, though, that Planck did not propose the photon. That would apparently have contradicted the very successful wave theory of electromagnetism and light, based on Maxwell’s equations. He just proposed the light was emitted in quanta by materials. Albert Einstein in 1905 proposed what we now call the photon to explain the photoelectric effect². Eq. (3.4) then gives the energy of the photon at some frequency f (or angular frequency ω).

Einstein later went on to extend the quantum theory of light in 1917, devising a particularly elegant derivation of Planck’s black-body spectrum while at the same time explaining the processes of emission and absorption of light by atoms. In particular, he deduced a key relation between absorption, spontaneous emission (the most common light emission, as in light bulbs) and the then-unknown process of stimulated emission (the process seen in lasers), a process he introduced for this argument. This 1917 work is known as “Einstein’s A and B coefficient argument”. We will return to the quantum mechanics of light and the black-body spectrum later.

Einstein’s proposal that light was actually made of quanta started the concept of “wave-particle” duality. This notion – that some entity could simultaneously have attributes of waves and of particles – seems to fly in the face of normal logic, and was not accepted or understood for some time; indeed, it remains one of the major conceptual challenges when starting to learn quantum mechanics³. It is, however, arguably not a problem in quantum mechanics if we avoid bringing along all the ideas of classical particles and classical waves. It is worth noting that wave-particle duality is a routine, everyday phenomenon in optical communications: We send light down optical fibers, treating it as a wave, but we generate and detect light using processes that are based on the ideas of photons, so we verify this idea literally trillions of times a day.

Quiz questions

- Q3.1.1 A black body at a temperature of 5800 K emits more light at 500 nm wavelength than it does at 1500 nm wavelength. True or False?
- Q3.1.2 A black body at a temperature of 3000K emits more light at 500 nm wavelength than it does at 700 nm wavelength. True or False?
- Q3.1.3 What is the correct photon energy in Joules for light of frequency 1015 Hz? (h in these suggested answers is Planck's constant, whose units are joule-seconds.)
- a) $h \times 10^{15}$

² It was actually for this proposal that Einstein was awarded the Nobel Prize in 1921.

³ And it is quite possible that it remains a conceptual challenge even after that!

- b) $\hbar \times 10^{15}$
- c) $h / 10^{15}$
- d) $2\pi h \times 10^{15}$

Problem

3.1.1 I need to generate as much electromagnetic radiation as possible in the (infrared) wavelength range beyond 1000nm (1 micron), but I only have a limited amount of electrical power to drive any light-bulb filaments. Should I drive a small number of light bulb filaments each using large fractions of my power and running at high temperature (e.g., 5800K) or should I use a large number of filaments each using small fractions of my power and running at a lower temperature (for example, 3000K). (You can presume that the filaments are black-body radiators, and that the only way the light bulb filament emits power is through electromagnetic “black-body” radiation – no heat is conducted away from the filaments by any other mechanism.) Justify your answer briefly.

3.2 The early quantum mechanics of atoms

Early experiments on atomic structure

Experiments on “cathode rays” in evacuated tubes, culminating in work by Joseph (“J. J.”) Thomson in 1897, showed that these rays were made up of very light particles (much lighter than an atom) with negative charge – what we now call electrons – that were components, in some way, of atoms. The structure of the atom itself was not clear however.

Ernest Rutherford in 1911, by scattering alpha particles using a gold foil, concluded that atoms had to have a very small charged nucleus, much smaller than the apparent overall size of the atom itself.

The Bohr model of the atom

Neils Bohr in 1913 made a key proposal, in what is known as the Bohr model of the atom. Bohr’s proposal was, first, that the electron in the hydrogen atom orbited round a positively charged nucleus, like a planet round the sun, with the Coulomb electrostatic attraction acting similarly to the gravitational attraction of the planet and the sun. But the second key aspect that Bohr proposed was that the angular momentum of the electron in its orbits should always be an integer multiple of “h bar” ($\hbar = h / 2\pi$). Such an orbit model is sketched in Fig. 3.5.

Thus Bohr was taking the radical step of using Planck’s constant, which had been introduced on a purely “ad hoc” basis by Planck to explain the black-body spectrum, and introducing it into the theory of matter. This model is successfully able to explain the basic energies of the hydrogen atom states, giving the formula

$$E_H(n) = -\frac{Ry}{n^2} \quad (3.5)$$

Here, n is an integer, starting at 1, and able to take any of the values 1, 2, 3, ... and so on. (n is now known as the “principal quantum number”.) Ry is a constant known as the Rydberg energy (or just “the Rydberg” for short). The “zero” of the energy here corresponds to the energy at which the hydrogen atom is just about to be ionized, which classically is equivalent to the “escape” energy, as in the minimum “escape velocity” of a space ship if it is to be able to escape the Earth’s gravitational pull.

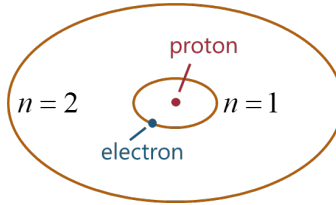


Fig. 3.5. Sketch of Bohr’s model of the hydrogen atom, with a point-like electron orbiting the proton like a moon round a planet, and restricted to orbits with angular momenta in units of $\hbar$.

It is relatively straightforward given Bohr’s assumptions to derive Eq. (3.5), coming up with the formula for the Rydberg energy in terms of fundamental constants, including the masses of the electron and the single proton that forms the nucleus of the hydrogen atom and, of course, Planck’s constant. We will return to that full expression later. For the moment, we note that the Rydberg’s value is

$$Ry \approx 13.6 \text{ eV} \quad (3.6)$$

Note that we have chosen to express the Rydberg energy in the energy unit of “electron-volts” or “eV” for short. The electron-volt is technically the amount of energy involved in moving an electron through an electrostatic potential difference of 1 volt (1 V). The relation between energy in electron-volts and energy in joules is

$$1 \text{ eV} \equiv |e| \text{ J} \approx 1.602 \, 176 \, 565 \times 10^{-19} \text{ J} \quad (3.7)$$

where e here is the magnitude of the charge of an electron in coulombs, that is

$$e \approx 1.602 \, 176 \, 565 \times 10^{-19} \text{ C} \quad (3.8)$$

and by $|e|$ we mean we are using the numerical value of e as the conversion factor here between these two systems of units.

Electron-volts are used extensively in quantum mechanics as units of energy because they are of a convenient size and they also have a simple physical meaning for particles whose charges are in integer units of the charge of the electron (which is the case for all atomic nuclei and all ions as well as the electron itself). When we are considering electronic and optoelectronic devices, electron-volts are particularly convenient units because there we are also often working with electrons and electrostatic potentials in volts.

We can use the formula Eq. (3.5) to explain the photon energies corresponding to the wavelengths of the spectral lines in the Balmer series. These particular spectral lines are then seen as corresponding to the various transitions from higher excited states down to the $n = 2$ level of the hydrogen atom, as shown in Fig. 3.1

To make the conversion between the wavelengths of the spectral lines and the transition energies, we note first that, quite generally for electromagnetic waves in a vacuum, the wavelength λ (in meters) is related to the frequency f (in Hz or cycles per second) by

$$\lambda = \frac{c}{f} \quad (3.9)$$

where c is the velocity of light in free space, a constant that in the modern approach to fundamental constants is simply chosen⁴ to have the definite value

$$c = 299\,792\,458 \text{ m/s} \quad (3.10)$$

If we follow through Bohr's approach here, we can also calculate the radius of the first orbit in that model (that is, for $n = 1$) which gives the parameter known as the Bohr radius a_0 with a value

$$a_0 \approx 5.29 \times 10^{-11} \text{ m} = 0.0529 \text{ nm} = 52.9 \text{ pm} = 0.529 \text{ \AA} \quad (3.11)$$

The unit “nm” is nanometers (10^{-9} m), the unit “pm” is picometers (10^{-12} m), and the unit “Å” is angstroms⁵ (10^{-10} m)⁶. In practice, the angstrom version may be the most widely used (picometers is comparatively rarely used in practice in this context), and it allows the convenient order-of-magnitude idea that a hydrogen atom is approximately about 1 Å in diameter. Though the modern model of the hydrogen atom is different from Bohr's approach here, this Bohr radius remains a useful characteristic size in discussing the hydrogen atom.

We should, however, be clear, without detracting from Bohr's remarkable advance here, that this model is not correct in several ways⁷. It was, nonetheless, a critical step in the

⁴ The meter is now defined as the length of the path traveled by light in vacuum during the time interval of $1/299\,792\,458$ of a second. The underlying reason for this way of approaching fundamental constants is that now we have a very precise frequency standard (and hence the time standard is defined as a specific number of cycles of our frequency standard) and so we use that time or frequency standard and this choice of the velocity of light to define the unit of length.

⁵ This unit is named after Anders Ångström (1841 – 1874), a pioneering Swedish spectroscopist. When the unit is written as a letter, it is always written with the ring diacritic, the small circle on top of the A, correctly corresponding to the first (Swedish) letter in his name. When written out as a unit, both the ring diacritic and the umlaut accent on the “o” are likely to be dropped completely, though it would be more correct to write ångström.

⁶ The angstrom is not technically one of the standard units in the SI system of units, but it is widely used because of its convenience in the quantum mechanics of atomic systems and for historical reasons.

⁷ Unfortunately, it has become a picture that non-scientists often view as the structure of the atom, with point-like electrons in circular orbits round a very small nucleus, a picture that is

advance of quantum mechanics, setting up the next phase of understanding of the quantum mechanics of the atom.

Perhaps the first puzzle of the Bohr model is that, because it suggests an orbiting electron, in a classical model of electromagnetism, it should be radiating electromagnetic waves all the time. In classical electromagnetism, accelerating a charge leads to radiated waves. We might not think the charge is accelerating in one sense here – in such a hypothetical circular orbit, the speed of the electron is not changing – but we must remember that acceleration is the rate of change of a vector quantity, the velocity. Changing the direction of the velocity vector in time is certainly also an acceleration. So there is at least an unresolved question here as to why these orbits are not radiating.

In fact, once we construct the correct model of the hydrogen atom using the Schrödinger equation, we find several other things “wrong” with the Bohr model. Though Bohr correctly proposed that angular momentum was quantized in units of $\hbar$, the lowest energy state actually corresponds to an angular momentum of zero, not 1, units. And, though higher energy states can indeed have larger angular momentum in units of $\hbar$, not all of them actually do, a point that again becomes clarified with the full Schrödinger equation solution.

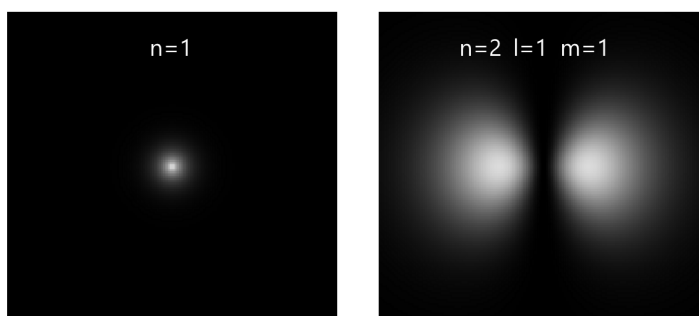


Fig. 3.6. Electron charge density, or equivalently the probability density of finding an electron at some position relative to the nucleus, in a hydrogen atom, shown here in cross-sections passing through the center of the atom for two different possible states.

More importantly, though, the main point that is not conceptually correct about the Bohr model, and one that could not really have been anticipated by Bohr at that time, is that the electron is not some point-like particle⁸ – some “dot” orbiting in these classical orbits. This is a deep and important point in quantum mechanics. When we are able later to construct the correct model using the Schrödinger equation, we find probability distributions or electron charge densities like those shown in Fig. 3.6. These are all “fuzzy clouds” – they are nothing like circular orbits of a point particle. For the lowest,

profoundly wrong. Bohr should not in any way be blamed for this sloppy but persistent misrepresentation, however! Bohr quite certainly accepted the later improved models, and continued to make major contributions to quantum mechanics.

⁸ Getting the idea of the electron as some sort of tiny point particle out of our heads is one of the major challenges in learning quantum mechanics!

$n=1$ state (also known as the 1s state), the distribution peaks in the middle, is spherically symmetric, and falls off smoothly. For the other state shown in Fig. 3.6, the distribution is more like a kind of “dumbbell” shape⁹; this one corresponds to one of the 2p orbitals of the hydrogen atom.

De Broglie’s hypothesis

The next key step in setting up the quantum mechanics of matter was the proposal by Louis de Broglie in 1924 that, just as there was a wavelength associated with light and light was also describable in terms of particles, then for particles of matter there should also be a wavelength associated with them. He proposed that wavelength λ was related to the magnitude of the momentum p of the particle. Specifically, he proposed

$$\lambda = \frac{h}{p} \quad (3.12)$$

an idea that can be fitted in with Bohr’s orbital idea by requiring integer numbers of wavelengths round an orbital. This proposal of matter as waves – a completely counter-intuitive one at that time – then suggests that there must be some wave equation describing these waves, which sets us up for Schrödinger’s wave equation.

Incidentally, Eq. (3.12) also turns out to be correct for the relation between the momentum of the photon¹⁰ and the corresponding wavelength λ of the electromagnetic wave, though we should be clear that the quantum mechanical wave for a particle like an electron is another kind of wave entirely.

Over the next 8 years or so, the core of the quantum mechanics that we use to understand ordinary matter and light was constructed by a collection of physicists and mathematicians, including Max Born, Paul Dirac, Werner Heisenberg, John von Neumann, Wolfgang Pauli, and Erwin Schrödinger.

Almost simultaneously, Heisenberg and Schrödinger proposed mathematical models, subsequently shown by Schrödinger to be equivalent, that could explain the hydrogen atom. Heisenberg’s approach (in 1925) is more abstract, being based on matrix algebra. On the face of it, Schrödinger’s wave equation (in 1926) is more tangible, giving a

⁹ To specify these other states of the hydrogen atom, we need to specify two other numbers, l – the orbital angular momentum quantum number, and m – the magnetic quantum number.

¹⁰ It might seem very strange that the photon should have momentum because it does not have any mass. However, we can actually deduce from ordinary classical electromagnetism that a light beam has momentum. The phenomenon of “radiation pressure” is well known – a light beam can indeed “push” an object quite measurably. If we consider the photons to have an energy hf for frequency f , then for a beam of a given intensity to have the radiation pressure that we get from classical electromagnetism, the photons would indeed have to have a momentum $p = h / \lambda$.

“wave” that we can visualize¹¹, and we will start below by taking this approach. In practice, we use both the matrix and wave approaches routinely today, moving smoothly between them. Actual calculations in quantum mechanics are quite likely to use matrices for example, even if we are thinking about wave functions.

Born, among other things, contributed the idea of how we interpret Schrödinger’s waves, and Pauli introduced the important idea of electron “spin” that is an additional sophistication we need to add on top of Schrödinger’s equation to model atoms with more than one electron. Dirac and von Neumann largely sorted out the mathematics of quantum mechanics, and Dirac also made other key contributions, such as a deeper understanding of the quantum mechanics of light, and the explanation of the origin of spin.

Quiz questions

- Q3.2.1 Choose the correct (approximate) value of the energy level of the hydrogen atom corresponding to $n = 1$ as given by the Bohr model: -13.6 eV, 13.6 eV, -13.6 J, or 13.6 J?
- Q3.2.2 The photon energy corresponding to the energy difference between the $n = 1$ and $n = 2$ energy levels is (in electron-volts) approximately which of the following: 10.2 eV, 10.2 J, 6.8 eV, or 6.8 J?
- Q3.2.3 The photon energy for light of wavelength 500 nm is approximately which of the following: 2.5 eV, 2.5 J, 2.1×10^{21} J or 2.1×10^{21} eV ?
- Q3.2.4 The Bohr model gives the correct picture for what an electron actually looks like in a hydrogen atom, with a point particle in circular orbits. True or False?
- Q3.2.5 Doubling the electron momentum (choose one answer)
- a) doubles its de Broglie wavelength λ
 - b) doubles its k value (note that $k = 2\pi / \lambda$)
 - c) multiplies its k value by 4
 - d) halves its k value

Problems

- 3.2.1 Calculate the wavelengths in nanometers of the spectral lines of the hydrogen atom corresponding to the transitions to the $n = 3$ energy level from each of the $n = 4, 5$, and 6 energy levels.
- 3.2.2* What would be the wavelength of the electromagnetic radiation whose photon energy is equal to 1 Rydberg (1 Ry)? (Such a photon energy would be required to raise an electron from the lowest level in the hydrogen atom to the point where

¹¹ In fact, as we will see, the “wave” in Schrödinger’s equation is quite a subtle concept, possibly without the “reality” that we normally ascribe to entities like water waves or electromagnetic waves.

it is just “ionized” – analogous to reaching escape velocity when launching a spacecraft from the earth).

- 3.2.3* An electron is accelerated from rest using an electrostatic potential of 3 V.
- What is the wavelength of the quantum mechanical wave corresponding to an electron with this much kinetic energy?
 - What is the (electromagnetic) wavelength of a photon with the same energy?
- 3.2.4 What is the de Broglie wavelength of each of the following objects, presuming in each case that we can treat them as a quantum mechanical particle of the given mass:
- A fullerene molecule (C_{60} , also called a “buckyball” because of its shape) of mass $m \approx 1.2 \times 10^{-24}$ kg at a speed $v = 200$ m/s?
 - A soccer ball of mass $m = 0.43$ kg moving at 25 m/s?
 - The Voyager space probe, $m \approx 722$ kg, moving at speed 17,260 m/s?
- 3.2.5 A red photon of wavelength 633 nm is reflecting directly backwards off a mirror in its path, changing its (photon) velocity from c to $-c$ in that direction. What is the momentum transferred to the mirror? What would this value be if there were 10^{10} red photons instead of one? (Note: The de Broglie formula also works for the momentum of a photon if we use the wavelength of the light as the wavelength λ in the formula. That is, the magnitude of the momentum of a photon is also $p = h / \lambda$.)

3.3 Schrödinger's wave equation

One point we need to note about quantum mechanics is that we cannot derive it from classical physics. We just have to postulate it. The only justification for some postulation is that it works – it correctly predicts what we observe in the world around us. So, for Schrödinger's equation, we simply have to postulate it also. Having said that, however, we can at least look for some starting points and ideas from the classical physics we already understand.

In our introduction here, we start with the simplest wave equation we know – the Helmholtz equation – which we will write here for what we call our “wavefunction” ψ . For a “monochromatic” wave in one direction z , of wavelength λ , we have

$$\frac{d^2\psi}{dz^2} = -k^2\psi \quad (3.13)$$

where

$$k = \frac{2\pi}{\lambda} \quad (3.14)$$

as defined before in the discussion of classical waves. As we know, at least before we start imposing boundary conditions, functions like $\sin kz$ and $\cos kz$ solutions of such

an equation (3.13), and any linear combination of them (that is, any sum of them with constant coefficients) is also.

In classical mechanics, quantities like wave amplitudes have to be real. Sometimes in dealing with classical systems we use complex numbers for mathematical convenience; complex exponentials¹² like $\exp(ikz)$ or, in time-dependent problems, $\exp(i\omega t)$, are often used as intermediate algebraic forms, but in the end we have to return to real wave amplitudes (typically, by adding in the complex conjugate at the end of the problem). Certainly from a mathematical point of view, the function $\exp(ikz)$ is a solution of Eq. (3.13) and, of course, from Euler's formula it is anyway a linear combination of $\sin kz$ and $\cos kz$

$$\exp(ikz) = \cos kz + i \sin kz \quad (3.15)$$

which would also qualify it as a solution to the linear differential equation (3.13).

In quantum mechanics, we make more use of the complex form, like Eq. (3.15). So our possible set of solutions to Eq. (3.13) includes all of $\sin kz$, $\cos kz$, and $\exp(ikz)$, as well as $\sin(-kz)$, $\cos(-kz)$, and $\exp(-ikz)$.

For the more general three-dimensional Helmholtz equation, we would now have

$$\nabla^2 \psi \equiv \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} = -k^2 \psi \quad (3.16)$$

where again we will be considering $\sin(\mathbf{k} \cdot \mathbf{r})$, $\cos(\mathbf{k} \cdot \mathbf{r})$, and $\exp(i\mathbf{k} \cdot \mathbf{r})$, as well as $\sin(-\mathbf{k} \cdot \mathbf{r})$, $\cos(-\mathbf{k} \cdot \mathbf{r})$, and $\exp(-i\mathbf{k} \cdot \mathbf{r})$.

Suppose, then, we now try to construct a wave equation for quantum mechanics, based on a Helmholtz equation, Eq. (3.13) and de Broglie's hypothesis, Eq. (3.12), which we now rewrite using Eq. (3.14) in terms of k as

$$k = \frac{2\pi p}{h} = \frac{p}{\hbar} \quad (3.17)$$

Incidentally, this gives another very common form for writing de Broglie's hypothesis, which is¹³

$$p = \hbar k \quad (3.18)$$

Hence we can rewrite our Helmholtz equation, Eq. (3.13), as

¹² We will use both of the notations $\exp(i\theta) \equiv e^{i\theta}$ here, with a preference for the former because it reduces the use of small superscripts. Note, incidentally, that we will use the physics notation of $i \equiv \sqrt{-1}$ instead of j , which is more common in engineering (possibly to avoid confusion with the use of i for current in engineering). Also, because in quantum mechanics we almost always write a time-oscillation in the form $\exp(-i\omega t)$ rather than the form $\exp(j\omega t)$ that is more common in engineering, sometimes we can think of i being $-j$, though this is not reliable.

¹³ This form $p = \hbar k$ is also correct for the photon momentum, where in that case the $k = 2\pi / \lambda$ is the wavevector magnitude corresponding to the electromagnetic wavelength λ .

$$\nabla^2\psi = -\frac{p^2}{\hbar^2}\psi \quad (3.19)$$

or

$$-\hbar^2\nabla^2\psi = p^2\psi \quad (3.20)$$

We presume we are interested in some particle of mass m (which we could presume for definiteness to be an electron, for example). So, we can divide both sides by the particle mass m to obtain

$$-\frac{\hbar^2}{2m}\nabla^2\psi = \frac{p^2}{2m}\psi \quad (3.21)$$

Now, we know from classical mechanics that

$$\frac{p^2}{2m} \equiv \text{the kinetic energy of the particle} \quad (3.22)$$

In general, in classical mechanics, we presume

$$\text{Total energy } (E) = \text{Kinetic energy} + \text{Potential energy } (V(\mathbf{r})) \quad (3.23)$$

So therefore we presume

$$\begin{aligned} \text{Kinetic energy} &= p^2 / 2m \\ &= \text{Total energy } (E) - \text{Potential energy } (V(\mathbf{r})) \end{aligned} \quad (3.24)$$

Therefore, we can propose to rewrite the Helmholtz equation, Eq. (3.21), as

$$-\frac{\hbar^2}{2m}\nabla^2\psi = (E - V(\mathbf{r}))\psi \quad (3.25)$$

or, with a minor rearrangement,

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r}) \right) \psi = E\psi \quad (3.26)$$

Eq. (3.26) is the Schrödinger equation (technically, the time-independent Schrödinger equation) for a particle of mass m .

Note, as we said at the beginning of this section, that we have not “derived” Schrödinger’s equation. We are merely suggesting it. We cannot derive it from first principles. The only justification for making this postulation of Schrödinger’s equation is that it works. And it does work. The first major triumph of Schrödinger’s equation¹⁴ is that it can explain the hydrogen atom.

¹⁴ The alternate and equivalent approach of the Heisenberg matrix model was also successfully used around the same time by Pauli to solve this hydrogen atom problem.

Quiz question

Q3.3.1 For each of the following equations, state whether the equation is a valid rearrangement of Schrodinger's (time-independent) equation.

a) $-\frac{\hbar^2}{2m}\nabla^2\psi = V(\mathbf{r})\psi$

b) $\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r})\right]\psi = E\psi$

c) $-\frac{\hbar^2}{2m}\nabla^2\psi = [E + V(\mathbf{r})]\psi$

d) $\nabla^2\psi = \frac{2m}{\hbar^2}[V(\mathbf{r}) - E]\psi$

Problems

3.3.1* In a region of space, a particle with mass m and with zero energy has a time-independent wavefunction in the z direction

$$\psi(z) = Az \exp\left(-\frac{z^2}{L^2}\right)$$

where A and L are constants. Determine the potential energy $V(z)$ of the particle.

3.3.2 Consider the Schrödinger equation for a particle of mass m with some specific total energy E in a uniform potential (i.e., the potential energy does not vary in space) of some specific value V . For simplicity here we consider the equation in just one dimension, i.e.,

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dz^2} + V\psi = E\psi$$

For such a situation, the wavefunction that solves this equation can take the form

$$\psi(z) = A \exp(ikz)$$

for some constant A and a wavevector magnitude $k = 2\pi/\lambda$ where λ is the corresponding wavelength of the electron wavefunction.

- (a) Find an expression for k in terms of E , V , m and fundamental constants.
- (b) If $E = 1.5$ eV, $V = 1$ eV and the particle is an electron, calculate the values of k and λ .

3.4 Probability density

Though we have now proposed a wave equation, we have to confront one of the subtler points in quantum mechanics, which is the “meaning” of the wavefunction. Of course, we might like to think it has the same simplicity of meaning as in classical waves; there, the wave amplitude itself corresponds to something directly measureable and meaningful. The amplitude could be an actual displacement, as in a wave on a string, or a pressure, as in acoustic waves in air, or at least something like the force on a charge, as in an electromagnetic wave. In quantum mechanics, however, it does not appear that the wavefunction has any such simple and direct meaning. Indeed, since in many cases we have to consider the quantum-mechanical wave amplitude to be a complex number, the fact that the wavefunction itself does not have to be considered as a measureable entity is actually convenient¹⁵.

It was Max Born who in 1926 proposed that the modulus squared of the wavefunction, should be interpreted as a probability density, $P(\mathbf{r})$, of finding the electron at¹⁶ the point $\mathbf{r}$, that is¹⁷

$$P(\mathbf{r}) \propto |\psi(\mathbf{r})|^2 \equiv \psi(\mathbf{r})^* \psi(\mathbf{r}) \quad (3.27)$$

The wavefunction itself, then, is not a probability or a probability density. It is, however, an example of what in quantum mechanics we can call a “probability amplitude” or a “quantum mechanical amplitude”. This concept of a probability amplitude ψ , as in Eq. (3.27), where we have to take the squared modulus to get a probability, is one that has no precedent in ordinary probability theory; it is one of the new mathematical ideas we have to introduce in quantum mechanics to have it explain the world around us.

With this additional postulation by Born, we have the quantum mechanical and mathematical basis for analyzing the hydrogen atom. We will postpone the solution of the hydrogen atom itself because it takes some time and effort to write down the necessary mathematics and solve the resulting equations, but we now have all of the physical postulates we need to do that.

3.5 The “particle in a box” model

A simple example quantum system, and one that turns out to be directly useful in analyzing actual modern semiconductor devices, is the so-called “particle in a box”. Here we imagine a “box” in an idealized and simple form. We consider only one

¹⁵ And possibly even something of a relief! How would we conceptualize a measureable yet complex entity?

¹⁶ Strictly, we should think of this as the probability of finding the electron within a small volume round about the point $\mathbf{r}$, divided by the magnitude of that small volume.

¹⁷ We express this formula with a “proportional to” sign (“ $\propto$ ”) because we have not yet dealt with the issue of making sure that all the probabilities add up to 1, a process called “normalization” that we can postpone for the moment.

direction, z , here for our waves for simplicity – what we could call a “one-dimensional” box. For definiteness, we could presume we are considering an electron, though this approach will work for any particle with some given mass, so we will just write a mass m . Now we can try to analyze this problem with our Schrödinger equation, Eq. (3.26).

Solving Schrödinger’s equation

Quantum mechanically, instead of having walls or posts to which to attach some “string”, we presume that we have infinitely high “potential barriers”, which we may also call infinitely high “walls”. We presume that the separation between these two walls is some distance L_z . This situation is illustrated in Fig. 3.7. We will sometimes call this an “infinite” or “infinitely deep” (potential) well.

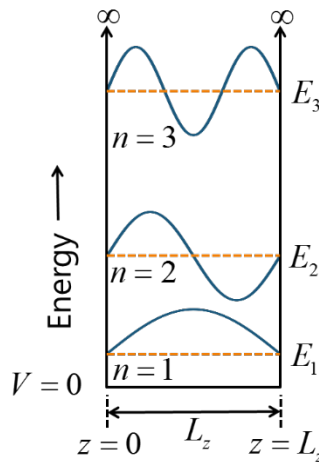


Fig. 3.7. Solutions for a particle in a one-dimensional box of width L_z and with infinitely high walls. Solutions are shown for the first three eigenenergies and the associated eigenfunctions, with the zeros for plotting the eigenfunctions chosen to be the dashed lines at each corresponding eigenenergy E_1 , E_2 , and E_3 here.

Formally, then, we can choose the potential energy V to be zero inside the box¹⁸. So, inside the potential well or box, because $V = 0$ inside the box, the Schrödinger equation Eq. (3.26) becomes

$$-\frac{\hbar^2}{2m} \frac{d^2\psi(z)}{dz^2} = E\psi(z) \quad (3.28)$$

or equivalently

$$\frac{d^2\psi(z)}{dz^2} = -\frac{2m_o E}{\hbar^2} \psi(z) \equiv -k^2 \psi(z) \quad (3.29)$$

where

¹⁸ It never matters where we choose the zero of energy to be – we can choose it anywhere we want. We simply have to be consistent in using whatever choice we make. It is often simplest just to choose the energy to be zero at some convenient point in our energy range of interest.

$$k^2 = \frac{2m_0 E}{\hbar^2} \quad (3.30)$$

We are presuming here that the potential energy rises to be arbitrarily (or infinitely) high at $z = 0$ and at $z = L_z$. Because we presume that any solutions will be for some finite energy E , we presume there is no possibility of finding the electron outside the box (that is, for $z < 0$ or $z > L_z$), and so we propose that the wavefunction ψ should be zero both at $z = 0$ and at $z = L_z$; that is, we have boundary conditions

$$\psi(0) = 0 \quad (3.31)$$

and

$$\psi(L_z) = 0 \quad (3.32)$$

Just as for the wave on a string, the general solution to Eq. (3.28) or Eq. (3.29) is

$$\psi(z) = A \sin(kz) + B \cos(kz) \quad (3.33)$$

where A and B are constants and, from Eq. (3.30)

$$k = \sqrt{\frac{2mE}{\hbar^2}} \quad (3.34)$$

The boundary condition, Eq. (3.31), means that we must choose $B = 0$ (we remember that $\cos(0) = 1$). So we are left with

$$\psi(z) = A \sin(kz) \quad (3.35)$$

The second boundary condition, Eq. (3.32), means that $kL_z = n\pi$ for some integer n , so

$$k = \frac{n\pi}{L_z} \quad (3.36)$$

But rewriting Eq. (3.34) gives

$$E = \frac{\hbar^2 k^2}{2m} \quad (3.37)$$

so the solutions of Schrödinger’s equation are, for the wavefunctions (eigenfunctions)

$$\psi_n(z) = A_n \sin\left(\frac{n\pi z}{L_z}\right) \quad (3.38)$$

and for the eigenvalues (eigenenergies)

$$E_n = \frac{\hbar^2}{2m} \left(\frac{n\pi}{L_z}\right)^2 \quad (3.39)$$

where, for the same reasons as in the case of waves on string (no purely-zero function, and only counting distinct functions once), we take only positive, non-zero values for n , that is, $n = 1, 2, 3, \dots$. These solutions are sketched in Fig. 3.7.

Normalization and probability

The wavefunctions as given by Eq. (3.38) are perfectly valid answers to this problem. With the wavefunction as stated in Eq. (3.38), the modulus squared of the wavefunction would be proportional to the probability, per unit length, of finding the electron “at” some point. We can be more specific about the probability density if we normalize the wavefunction, that is, if we arrange that these probabilities all add up to one in some useful sense.

Using the wavefunction as written in Eq. (3.38), we can integrate the modulus squared over the width of the potential well – which is the only range of positions for which the wavefunction is non-zero – so we can add up all of these probability densities. Specifically, then, we would have¹⁹

$$\int_0^{L_z} |\psi_n(z)|^2 dz = \int_0^{L_z} |A_n|^2 \sin^2\left(\frac{n\pi z}{L_z}\right) dz = |A_n|^2 \frac{L_z}{2} \quad (3.40)$$

So, if we choose

$$|A_n| = \sqrt{\frac{2}{L_z}} \quad (3.41)$$

then the answer to this integral is 1, as desired, that is,

$$\int_0^{L_z} |\psi_n(z)|^2 dz = 1 \quad (3.42)$$

Conventionally, we choose A_n to be real for simplicity²⁰, so we can write the corresponding wavefunction solutions as

$$\psi_n(z) = \sqrt{\frac{2}{L_z}} \sin\left(\frac{n\pi z}{L_z}\right) \quad (3.43)$$

¹⁹ This particular integral is particularly simple because it is over complete periods of the sine squared function, and the average of the sine squared function is $\frac{1}{2}$ over any complete period.

²⁰ We do not have to choose a real coefficient. We can multiply any eigenfunction by a complex number of unit magnitude and it will still be an eigenfunction with the same normalization.

The form as in Eq. (3.43) is then the normalized version of the wavefunctions (or eigenfunction)²¹. The normalized wavefunctions here have dimensions of one over the square root of distance²².

We can plot out the modulus squared of the wavefunctions, as shown in Fig. 3.8. When we use the normalized version of the wavefunctions, the area under each one of these plots, relative to its corresponding zero line, is 1.

To evaluate the probability of finding a particle in a given region of space, then we formally need to integrate the probability density over that region. Suppose, for example, that we knew that the particle (for example, an electron) was in state $n = 1$ in such a box, and we wanted to know the probability P_s of finding the electron between some point z_1 and some other point z_2 . Then, using the normalized wavefunction from Eq. (3.43), we simply need to integrate the modulus squared of the normalized wavefunction over these limits.

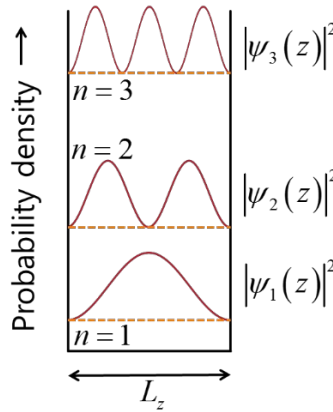


Fig. 3.8. Plot of the modulus squared of the wavefunctions for the first three levels of the particle in a box, with the corresponding horizontal dashed lines being the zeros for each different plot.

The answer²³ would be

²¹ Note, incidentally, that here the normalization constant A_n happens to be the same for all choices of n , though generally in quantum mechanical problems such normalization constants are likely to be different for different eigenfunctions.

²² For the wavefunctions corresponding to three dimensional problems, we similarly end up with normalized wavefunctions having the dimensions of one over the square root of volume.

²³ We are using the trigonometric identity $\sin^2(\alpha) = (1/2)[1 - \cos(2\alpha)]$ here.

$$\begin{aligned}
P_s &= \int_{z_1}^{z_2} \left(\sqrt{\frac{2}{L_z}} \right)^2 \sin^2 \left(\frac{\pi z}{L_z} \right) dz = \frac{2}{L_z} \int_{z_1}^{z_2} \frac{1}{2} \left[1 - \cos \left(\frac{2\pi z}{L_z} \right) \right] dz \\
&= \frac{1}{L_z} \int_{z_1}^{z_2} 1 dz - \frac{1}{L_z} \int_{z_1}^{z_2} \cos \left(\frac{2\pi z}{L_z} \right) dz = \frac{1}{L_z} [z]_{z_1}^{z_2} - \frac{1}{L_z} \frac{L_z}{2\pi} \left[\sin \left(\frac{2\pi z}{L_z} \right) \right]_{z_1}^{z_2} \\
&= \frac{(z_2 - z_1)}{L_z} - \frac{1}{2\pi} \left[\sin \left(\frac{2\pi z_2}{L_z} \right) - \sin \left(\frac{2\pi z_1}{L_z} \right) \right]
\end{aligned} \tag{3.44}$$

P_s is then the shaded area in Fig. 3.9.

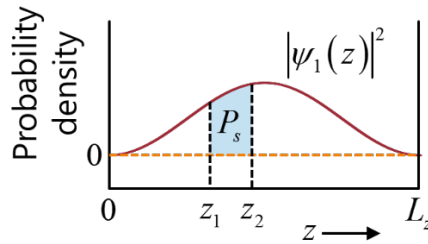


Fig. 3.9. Illustration of the area, P_s , under the probability density curve for the first state of a particle in a box, corresponding to the probability of finding the particle between z_1 and z_2 .

Note that this result in Eq. (3.44) is dimensionless – it is just a number, as we would expect for a probability. Note also that, if we choose $z_1 = 0$ and $z_2 = L_z$ then this expression, Eq. (3.44), correctly gives us the answer 1 – the particle must be somewhere inside the box.

Nature of the particle-in-a-box solutions 6d

Though this particle-in-a-box problem is a relatively simple one mathematically, it shows many of the general features of the solutions to quantum mechanical problems using Schrödinger's equation, Eq. (3.26).

Energy eigenvalues, eigenfunctions, and eigenstates

First, we see that we only have solutions to this equation for very specific energies, which are the eigenvalues as given by Eq. (3.39), with associated eigenfunctions, as given by Eq. (3.43). As we have seen, the idea of eigenvalues and eigenfunctions is by no means just one that comes up in quantum mechanics. Many classical problems of oscillators and waves show such behavior. The difference here is that we are seeing this behavior when considering matter. This behavior arises here because we are also considering matter to be describable by waves. In quantum mechanics, these solutions of the (time-independent) Schrödinger equation are called energy eigen solutions, with energy eigenvalues (or eigenenergies) and energy eigenfunctions.

Compared to the previous classical world, we are now seeing that, at least in this example problem, asking for solutions for which there is a definite energy E associated

with the solution leads to the conclusion that there are only very specific, discrete values of that energy that are possible²⁴. That behavior is different from prior classical models of matter. We will also refer to such solutions as energy eigenstates; these are solutions that are telling us possible states of matter (here, our particle) for this problem. These discrete states can also be indexed with a “quantum number” – here the integer n . We regard the “state” here as having these properties of an eigenenergy, an eigenfunction, and an appropriate quantum number²⁵ to index it.

Parity of eigenfunctions

One mathematical property of the eigenfunction solutions to this problem is what is called “parity”. Because this problem is rather symmetric (the potential energy is symmetric about the center of the well), the solutions also are influenced by that. In particular, with respect to the center of the well, the eigenfunctions are either (i) “symmetric” or have what is called “even parity”, which means the value of the function some distance to the right of the center is the same as the value of the function the same distance to the left of the center, or (ii) they are “antisymmetric” or have “odd” parity, which means the value of the function some distance to the right of the center is minus the value of the function the same distance to the left of the center. Often for brevity, we call functions with “even” parity just “even” functions, and similarly those with “odd” parity just “odd” functions.

For the eigenfunctions here, relative to the center of the potential well, as we can see from Fig. 3.7, the solutions with $n = 1, 3, 5, \dots$ have “even” parity, and those with $n = 2, 4, 6, \dots$ have “odd” parity. (Note that the parity of the functions does not happen to correspond with whether the quantum numbers themselves here are even or odd, so the odd-ness or even-ness of the quantum numbers is not itself a good guide to the parity of the functions.)

Not all quantum mechanical problems will have solutions with these “definite” parities (that is, either odd or even). Potentials that are not symmetric, such as a “tilted” potential well in which the bottom of the well is sloping in one direction or the other, can have solutions that are neither symmetric nor antisymmetric, and so those solutions can be said not to have definite parity.

Problems such as the hydrogen atom, which is a problem that is symmetric in many ways, often do have solutions with quite definite parities, as well as having other symmetries, such a rotational symmetries in which rotating by some angle or angles around some axis leaves the eigenfunctions unchanged. Symmetries generally can be very useful in simplifying solutions of these and other wave problems, and parity is one of the simplest examples of symmetries that such eigenfunctions may possess.

²⁴ In quantum mechanics in problems without confining boundaries, or for energies above the “height” of bounding potentials, it is also possible to have continuous ranges of eigenenergies that are not restricted to discrete values, but for problems where we can at least imagine the solutions are bounded by some box, we will have discrete energy eigenvalues.

²⁵ More generally, in more complicated problems, such as the hydrogen atom, we may end up with a set of quantum numbers to index the state.

Another obvious feature of the successive eigenfunctions in this particle-in-a-box problem is that each one has one more “zero” in it than the preceding one. For example, the $n = 1$ solution has no zeros (other than the ones at the walls), the $n = 2$ solution has a zero in the middle of the well, the $n = 3$ solution has yet one more zero, and so on. This also is common mathematical behavior and helps to ensure the orthogonality of different eigenfunction solutions.

Quantum confinement

This quantum mechanical analysis of a particle in a box is showing behaviors that are quite different from our notions of what would happen classically to a particle in a box in at least three different ways, all of which we can regard as being consequences of what we call “quantum confinement” in the box; by quantum confinement we mean that there is a set of quantum phenomena that emerge if we confine some system (here, a particle) on a small scale.

Discrete energy levels

There are only specific energies for which there are solutions to this equation. Classically, there would be no such restriction, nor even any basis for it, for a particle in a box. Even if we regarded the particle as being in some oscillatory classical state, bouncing back and forwards between the walls, the particle could have any kinetic energy; more kinetic energy would just mean it would bounce faster back and forwards in the box, but there would be no “discretization” of the frequencies at which it could do this, or of the amount of kinetic energy it could possess.

Zero-point energy

For a given particle in a given size of box, there is a minimum energy the particle can have, which is often known as the “zero-point energy”. Strangely enough, quantum mechanically, the particle can never simply have the energy corresponding to “sitting at the bottom of the box”, but must at least have some energy larger than that. In our particle-in-a-box model, this zero-point energy corresponds to the energy of the $n = 1$ state. Again, classically, there is no precedent for this; in a classical view, of course a particle can just “sit still” at the bottom of the box, with no kinetic energy, but in quantum mechanics, apparently, in general, it cannot do that.

This zero-point energy becomes larger as we make the box smaller; indeed, it can get larger quadratically as we reduce the size scale of the box, as we can see directly from the expression Eq. (3.39).

Non-uniform distribution

No matter what state we put the particle in, it is not at all uniformly distributed over the box. For any of the eigenstates, the probability distribution is in one of a set of “standing wave” patterns, as we can see in Fig. 3.8. Again, there is no precedent in classical mechanics for this kind of behavior for a particle. There we might again just imagine the particle bouncing back and forward from one wall to the other, which would give it an essentially uniform probability distribution. Even more perplexing is the notion that there are places inside the box, and not just at the walls, where in specific states there is no possibility of finding the particle; these places coincide with the zeros of the

eigenfunctions. This phenomenon is difficult to conceptualize if one takes the classical notion that the particle would have to “pass through” that point to get to the other side; if, however, we discard any need for that classical notion by discarding the idea of a “point” particle that has to have a definite position, there is arguably no particular problem in quantum mechanics with this phenomenon.

Order of magnitude of sizes and energies

A simple calculation shows the size scales and energies at which we expect to see these consequences of quantum confinement. Suppose, for example, we confine an electron in a 5 Å (0.5 nm) thick “box”. The electron has a mass

$$m_e \approx 9.109\,382\,91 \times 10^{-31} \text{ kg} \quad (3.45)$$

Using the expression, Eq. (3.39), and calculating the first allowed level for the electron – that is, for $n = 1$ – leads to a “zero-point energy” of

$$E_1 = \frac{\hbar^2}{2m_e} \left(\frac{\pi}{5 \times 10^{-10}} \right)^2 \approx 2.4 \times 10^{-19} \text{ J} \approx 1.5 \text{ eV} \quad (3.46)$$

(relative to the “bottom” of the well) and a separation between the first and second energy levels of

$$E_2 - E_1 \equiv 3E_1 \approx 4.5 \text{ eV} \quad (3.47)$$

These energies of small numbers of electron-volts are typical of major energy separations between the levels of atoms, which generally have sizes on roughly this angstrom scale. (Photons of visible light also have energies of the order of a couple of electron-volts, so their interaction with atoms also is of a quantum mechanical nature, as we might expect.)

Hence we see that even this very simple model is beginning to give us the scales of actual quantum mechanical phenomena. Generally, we expect quantum mechanical effects for matter to become stronger as we make things smaller. As a general sense of scale, the properties of materials are roughly independent of the size of the piece of material for scales of microns or larger. If we look carefully at the properties of objects on scale of about 100 nm or so, we can begin to see quantum confinement effects. For objects at a size scale of about 10 nm, quantum confinement effects are quite clear. Indeed, they are used routinely with semiconductor materials, for example, to tune their optical properties solely based on the physical dimensions.

By the time we get to the single nanometer scale, quantum effects are dominating the properties of a piece of material, drastically changing its properties compared to normal “bulk” properties. Quantum confinement from sizes on the single nanometer scale can be used to change the color of microcrystallites of semiconductor materials, for example. Such nanometer size scales are now routine for various nanotechnologies, including the fabrication methods used for transistors in silicon chips, for example, and the layered semiconductor structures used for modern optoelectronic devices like semiconductor lasers and light-emitting diodes.

Simply put, we cannot begin to understand the properties and operating principles of modern devices, especially those based on nanotechnology, without this physics of quantum confinement.

Quiz questions

Q3.5.1 Consider a particle of mass m in an infinitely deep potential well of thickness L_z , and with potential energy chosen to be $V = 0$. For each of the following, state whether it is a possible value of the energy E .

a) $\frac{\pi^2 \hbar^2}{2mL_z^2}$

b) $\frac{2\pi^2 \hbar^2}{2mL_z^2}$

c) $\frac{3\pi^2 \hbar^2}{2mL_z^2}$

d) $\frac{4\pi^2 \hbar^2}{2mL_z^2}$

Q3.5.2 Consider a particle of mass m in an infinitely deep potential well of thickness L_z with position running from $z = 0$ to $z = L_z$. For each of the following, state whether it is a possible wavefunction for this infinitely deep well.

a) $\sqrt{\frac{2}{L_z}} \sin\left(\frac{\pi z}{L_z}\right)$

b) $\sqrt{\frac{2}{L_z}} \cos\left(\frac{\pi z}{L_z}\right)$

c) $\sqrt{\frac{2}{L_z}} \sin\left(\frac{3\pi z}{2L_z}\right)$

d) $\sqrt{\frac{2}{L_z}} \cos\left(\frac{2\pi z}{L_z}\right)$

Q3.5.3 For "particle in a box" wavefunctions for a (1-dimensional) box of thickness L_z with position running from $z = 0$ to $z = L_z$, state whether each of the following are normalized eigenfunctions.

a) $\sqrt{\frac{L_z}{2}} \sin\left(\frac{3\pi z}{L_z}\right)$

b) $\sqrt{\frac{2}{L_z}} \sin\left(\frac{4\pi z}{L_z}\right)$

$$\text{c) } \sin\left(\frac{\pi z}{L_z}\right)$$

$$\text{d) } \frac{2}{L_z} \sin\left(\frac{8\pi z}{L_z}\right)$$

Problems

- 3.5.1* Consider an electron in a one-dimensional potential well with infinitely high potential barriers on each side and of thickness $L_z = 1.5$ nm.
- What is the zero-point energy of the electron in this potential well, relative to the energy of the bottom of the well? (Give your answer in electron-volts)
 - What is the energy separation of the first and second energy eigenstates for the electron in this well? (Give your answer in electron-volts)
- 3.5.2 Suppose we have an electron initially in free space, with no confining barriers of any kind. Now suppose we confine it in one direction with a pair of parallel “walls” that are effectively infinitely high (i.e., correspond to a potential energy that is effectively infinitely large). How far apart should the walls be if we want to give the electron a zero-point energy (i.e., a lowest possible energy, compared to the “free” electron’s minimum energy) of each of the following amounts?
- 25 meV (which, incidentally, is approximately the magnitude of the thermal energy at room temperature)
 - the same energy as a photon that corresponds to the electromagnetic wavelength used in long-distance optical fibers (which is about 1.5 μm in air or vacuum)
 - the same energy as the separation between the $n = 2$ and $n = 3$ levels of the hydrogen atom
- 3.5.3 An electron is confined in an infinite square well of width L .
- Calculate the energy separation in eV of the following states.
 - $L = 10$ nm, $n = 3 \rightarrow n = 1$
 - $L = 5$ nm, $n = 5 \rightarrow n = 3$
 - Suppose in each case that the system could emit a photon by the electron “falling” from the upper state of the pair to the lower one, with a photon energy corresponding to the separation between the upper and lower states. In each case, what would be the wavelength of the emitted photon? (Note: the energy of a photon is $E = \hbar\omega = hf$, where ω is the angular frequency and f is the frequency.)
- 3.5.4 For each of the following proposed functions, state whether it is an energy eigenfunction solution for Schrödinger’s time-independent equation for a particle-in-a-box problem for a box of thickness L_z , presuming infinitely high “walls” on

either side. Presume $z = 0$ at the “left” wall of the box and $z = L_z$ at the “right” wall of the box. [Note: any such solutions are not required to be normalized.]

(i) $\cos\left(\frac{\pi z}{L_z}\right)$

(ii) $\sin\left(\frac{\pi z}{2L_z}\right)$

(iii) $16\sin\left(\frac{5\pi z}{L_z}\right)$

(iv) $2\sin\left(\frac{\pi z}{L_z}\right) + 3\sin\left(\frac{3\pi z}{L_z}\right)$

(v) $2\exp\left(\frac{2i\pi z}{L_z}\right) - 2\exp\left(-\frac{2i\pi z}{L_z}\right)$

3.5.5 In the semiconductor GaAs, electrons behave as if they have a mass that is approximately 0.07 times the mass of the free electron. Suppose we have a thin layer of GaAs, of thickness L_z to be determined, that we will treat as if it was a one-dimensional potential well with infinitely high walls on either side. We want to make a device that emits light at an infrared wavelength of $10\text{ }\mu\text{m}$ by having electrons transition from the $n = 2$ level in this well to the $n = 1$ level. How thick should we make the layer (that is, what value should we choose for L_z)?

3.5.6* Find the normalization factor A for the following functions such that

$$\int_{-\infty}^{\infty} |\psi(z)|^2 dz = 1.$$

a) $\psi(z) = A\exp(-ikz)$ for $-L < z < L$.

b) $\psi(z) = A\exp(-\kappa|z|)$ for $-\infty < z < \infty$. (Presume $\kappa > 0$.)

c) $\psi(z) = A\cos\left(\frac{2\pi z}{L}\right)$ for $-\frac{L}{4} < z < \frac{L}{4}$.

3.5.7* For a particle that has each of the following normalized wavefunctions $\psi(z)$ in the z direction, calculate the probability of finding a particle in the given regions.

a) $\psi(z) = \sqrt{\frac{2}{L}} \sin\left(\frac{3\pi z}{L}\right)$ for $\frac{L}{3} < z < \frac{2L}{3}$.

b) $\psi(z) = \sqrt{\frac{2}{L}} \sin\left(\frac{2\pi z}{L}\right)$ for $0.8L < z < L$.

3.5.8 Consider an electron in an infinitely deep potential well in the second energy eigenstate of that well (i.e., $n = 2$).

- (i) Presuming the well has a thickness L_z , give an expression (with any integrations completed) for the probability of finding the electron between a position z_1 and a position z_2 in the well.

Presume now that $L_z = 8$ nm, for the electron, give the probability that we would find the electron in the each of the following regions of the well.

- (ii) Between 0 and 4 nm from the left of the well.
 (iii) Between 1 and 2 nm from the left of the well.
 (iv) Between 3 and 5 nm from the left of the well (i.e., in the middle 2 nm of the well).
 (v) Between 3.9 and 4.1 nm from the left of the well.

3.6 Waves and measurement

De Broglie's hypothesis and Schrödinger's wave equation were both proposed before there was any direct evidence of the wave nature of electrons. In 1927, however, clear experimental evidence emerged from the observation of diffraction of electrons, from the surface of nickel by Clinton Davisson and Lester Germer and with gold films by George Thomson²⁶. Diffraction of electrons by crystal surfaces is a standard diagnostic technique today in crystal growth. Modern electron microscopes, following initial successful development in the 1930s, routinely exploit the short wavelengths of accelerated electrons to obtain images with resolutions well beyond those of conventional optical microscopes, down even to the size scales of atoms themselves.

Examining diffraction phenomena for electrons also allows us conceptually to explore more of the consequences of this quantum mechanical view of matter and of the foundations of quantum mechanics, and we consider some of these consequences next.

An uncertainty principle

One well-known phenomenon seen with classical waves is diffraction from an aperture. This is illustrated in Fig. 3.10, which shows calculated wave patterns resulting from plane waves on the left, of wavelength λ , being incident on different apertures in a mask. When we pass waves through a small aperture, we see quite clearly at large distances that the resulting waves spread out. If we choose smaller apertures, the spreading angle is larger.

For an aperture of size d , the full angle between the minimum in the wave pattern at the top and at the bottom is approximately $\sim \lambda / d$. A more characteristic angle given some rough half-width of this angular spread would be significantly smaller than this, and it

²⁶ George Thomson was the son of "J. J." Thomson, who had proposed the electron as a particle in 1897. Both father and son separately won Nobel Prizes, George Thomson's being shared with Clinton Davisson for this work.

will be convenient for our algebra here²⁷ to take this characteristic half-width angle to be approximately

$$\theta_c \sim \frac{\lambda}{2\pi d} \quad (3.48)$$

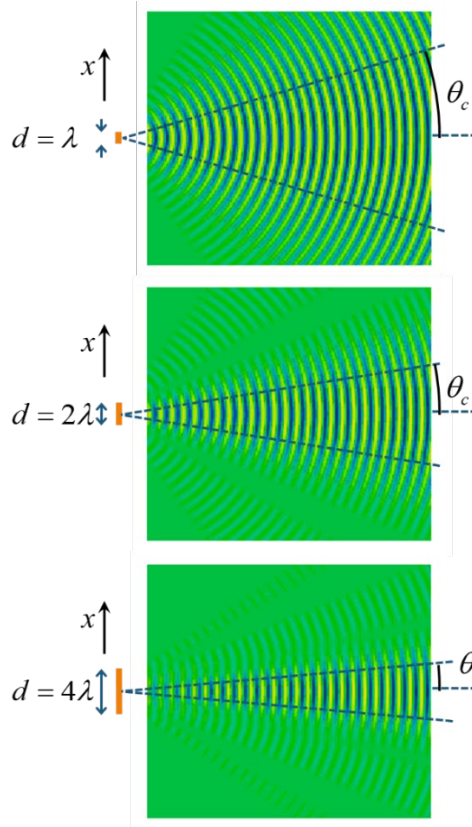


Fig. 3.10. Diffraction of monochromatic waves of wavelength λ , from apertures of widths $d = \lambda$, 2λ , and 4λ , showing characteristic diffraction angles θ_c for each d .

Note in this behavior that the better we know the position of the “source” of the waves in the aperture, as given by the size of the aperture, the less well we know the angle of the waves, or at least there is a larger spread in angle the waves. Relations like these occur routinely with waves in classical physics, and they represent classical “uncertainty principles”, where we have a reciprocal relation between the uncertainty of one quantity and the uncertainty in another. The relation between the uncertainty in the angular direction of the wave, as given by the diffraction angle, and the uncertainty

²⁷ The overall form of this diffraction angle formula is quite valid, at least for small angles. Our choice here of a factor of 2π in the bottom line is a reasonable number to choose to give a characteristic diffraction angle, but, honestly, this specific choice is just so we get to the “right” answer at the end! That right answer can be justified with more rigorous definitions and mathematics, but we are not giving either of those here.

in the position of the wave, which is at least the distance corresponding to the width of the aperture, is a good classical example of an uncertainty principle.

Another common kind of relation is one encountered with signals in time. If we have a very short “pulse” of a signal of width $\sim \Delta t$, then we know the timing of the signal well because we can know with this degree of precision at what time the pulse was emitted; but we will find that it is not meaningful to ask for the precise “frequency” of the pulse because it would necessarily be made up out of a range of frequencies; we could write that uncertainty in the frequency in angular frequency terms as $\Delta\omega$.

Fourier analysis can formally relate these uncertainties in timing and frequencies. With appropriate definitions in Fourier analysis, we can mathematically derive a relation like

$$\Delta\omega\Delta t \geq \frac{1}{2} \quad (3.49)$$

which tells us that a short pulse has to be viewed as being made up out of a large range of frequencies, and conversely, if the frequency range in some signal is small, then it must correspond to a long pulse. It is always possible to do worse than the minimum product – for a given frequency range, we can have an even longer pulse or signal for example – hence the inequality in Eq. (3.49).

Now if we think of quantum mechanics and particles, we can think of the width of the aperture as telling us the uncertainty we have in the position in the x direction of a photon if it is an electromagnetic wave, or of an electron if it is a quantum mechanical wave. If we think of this uncertainty in position as being $\pm \Delta x$ relative to the midpoint, then we can take $\Delta x = d / 2$.

For a wavelength λ , the magnitude of the wavevector is $k = 2\pi / \lambda$. If we consider a wavevector oriented at up to about some angle θ_c relative to the horizontal, then we can consider it to have a component of the wavevector in the x direction up to about $\Delta k_x = k \sin \theta_c \approx k \theta_c$, at least for small angles θ . Hence, using Eq. (3.48) for θ_c , we have

$$\Delta k_x \approx k \theta_c \geq \frac{2\pi}{\lambda} \frac{\lambda}{2\pi} \frac{1}{2\Delta x} = \frac{1}{2\Delta x} \quad (\text{or, equivalently, } \Delta k_x \Delta x \geq \frac{1}{2}) \quad (3.50)$$

We have written Eq. (3.50) with a “ $\geq$ ” sign rather than just a “ $\approx$ ” because it is always possible to have beams that diverge by more than this minimum characteristic angle. For example, if we start out with a beam that is already expanding, even a large beam can have a large divergence angle; the argument we used to illustrate these diffraction angles was for the “best” case where we start out with a “plane” wave that is not initially expanding. We can note immediately the similarity between these two “classical” uncertainty principles, Eqs. (3.49) and (3.50).

At least for electrons²⁸, from de Broglie’s hypothesis we know that the momentum $p = \hbar k$, so we take the uncertainty Δp_x in the momentum in the x direction to correspond with the uncertainty Δk_x we have in the component of the wavevector in the x direction – that is, we interpret $\Delta p_x \equiv \hbar \Delta k_x$. Hence, from Eq. (3.50), we have

²⁸ In fact, this same relation also is valid for photon momentum.

$$\Delta p_x \Delta x \geq \frac{\hbar}{2} \quad (3.51)$$

where we have interpreted our calculations here to give a minimum estimate of the divergence of the beam or, equivalently, the uncertainty in the momentum in the x direction. It is certainly possible to have larger divergence angles than those shown here, for example if the incident wave in the aperture is not a plane wave or has some structure to it.

Eq. (3.51) represents an uncertainty principle based on this rather informal derivation from diffraction angles. In fact, because of our choices of numbers here (which are admittedly somewhat arbitrary), this result corresponds to Heisenberg's uncertainty principle for position and momentum that can be derived with more precise definitions and mathematics.

One important point to make here is that there really is nothing surprising about the existence of uncertainty principles for waves or for signals in time. They are a routine part of the classical world. For a classical wave, it is meaningless to ask to know simultaneously both the "position" of a wave and its direction. If it has a precise direction, then it is a plane wave, and it is arbitrarily wide. If it has a precise position, then it is a wave at some point source, and it will spread out over a wide angle. With careful mathematical definitions and some more sophisticated mathematical proofs (which are formally the same in both classical and quantum mechanical worlds), we get uncertainty principles that are mathematically identical in form to the quantum mechanical one of Eq. (3.51).

The uncertainty principle and measurement

There are two aspects that distinguish the quantum mechanical uncertainty principle as in Eq. (3.51) from our previous classical thinking. The first is just the unexpected notion that we can describe quantum mechanical particles as waves. Given that we can, however, then uncertainty principles follow for the same reasons in the classical world. Viewed this way, there is nothing bizarre or mysterious in saying that a "particle" cannot simultaneously have a well-defined position and a well-defined momentum; this is a natural consequence of the particle also having wave character, and this kind of uncertainty relation happens for waves in the classical world also.

At the time when Heisenberg's uncertainty principle was being introduced, there was, however, a good deal of confusion about the idea of quantum mechanical measurement; in the original discussion, this idea of measurement is also mixed in with these notions of a particle not simultaneously having well-defined position and momentum. Popular discussions of Heisenberg's uncertainty principle often state it as saying that the act of measuring something changes it. There are, however, two quite separate ideas here, and a modern view of this would be to make that separation quite clear.

So, we have uncertainty principles like Eq. (3.51) that should be no surprise to us given the wave character we are now presuming for matter, which say it is meaningless for us to try to require a particle simultaneously to have a well-defined position and a well-defined momentum; that follows naturally from a wave picture.

The second, and much more difficult, issue – namely, that measuring something changes it – is really quite a different one that is better viewed as being something quite distinct from the uncertainty principle as we have been discussing it²⁹. That issue is known in quantum mechanics as the “measurement problem” and it is associated with the “collapse of the wavefunction”.

Diffraction by two slits

Young’s slits 7b

In the early 1800s, Thomas Young performed a classic experiment that showed that light had to be viewed as a wave. This is the “two-slit” experiment. This experiment also allows a measurement of the wavelength of light without requiring any ruler on the scale of the wavelength of the light itself. Our interest here is mostly in the conceptual aspects of this kind of experiment when we are considering wave-particle duality and the measurement problem in quantum mechanics, but to lay the groundwork for that discussion, we start by explaining Young’s slits for waves in general.

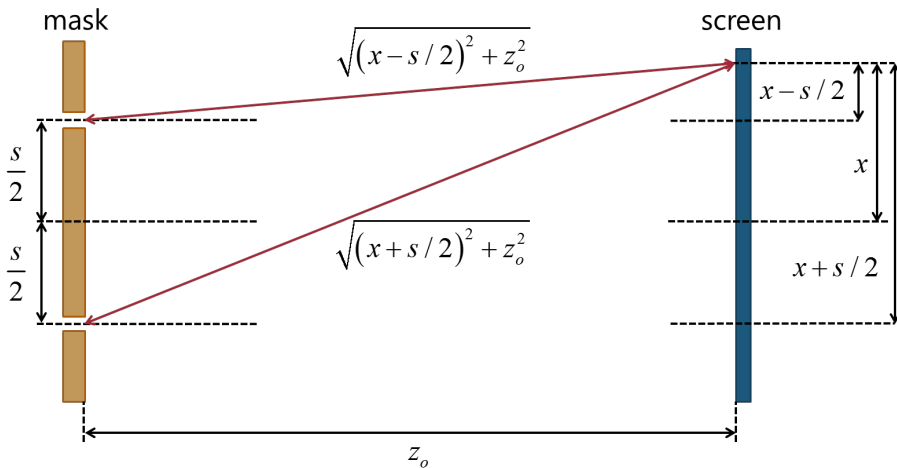


Fig. 3.11. Geometry of Young’s slits experiment (not to scale), with narrow slits a small distance s apart at a distance z_o from a screen.

The layout for this experiment is sketched in Fig. 3.11. A mask has two slits cut in it, separated by a distance s . Plane waves will be incident from the left. In the spirit of Huygens’ principle, the slits will effectively behave like point sources of light. These two sources will be in phase with one another because they are derived from the same plane wave, whose phase fronts are presumed to be vertical. A screen is placed some large distance z_o to the right. We will be interested in some point on the screen at a distance x from the midpoint. In an actual experiment, the slits will be quite close to one

²⁹ These two issues of the uncertainty principle and the measurement problem are linked if we view the Δp_x and Δx as being statistical standard deviations we deduce from repeated experiments measuring these quantities, but they are not linked if we think of these quantities as just being attributes of wave motion.

another, for example only a moderate number of wavelengths apart. The distance z_o will be much larger than s , so the diagram in Fig. 3.11 is not to scale; the slits will be relatively much closer and narrower, and the distances z_o and likely also x will be very much larger than s , though x should still be only a small fraction of z_o .

To get from the two different slits to some specific point x on the screen, the waves will travel different distances. This will lead to interference on the screen. We can understand that interference pattern by looking at the differences in the distances.

The distance d_1 from the upper slit to this point x on the screen will be, by Pythagoras' theorem,

$$d_1 = \sqrt{\left(x - \frac{s}{2}\right)^2 + z_o^2} = z_o \sqrt{1 + \left(\frac{x - s/2}{z_o}\right)^2} \quad (3.52)$$

Since we are presuming x and s are much smaller than z_o , we can approximate the square root by its lowest-order power series expansion – that is, for small a , $\sqrt{1+a} \approx 1 + a/2$ – so we can write, approximately,

$$d_1 \approx z_o + \frac{\left(x - s/2\right)^2}{2z_o} = z_o + \frac{x^2}{2z_o} + \frac{s^2}{8z_o} - \frac{sx}{2z_o} \quad (3.53)$$

Similarly, the distance d_2 from the lower slit to the point x on the screen will be

$$d_2 = \sqrt{\left(x + \frac{s}{2}\right)^2 + z_o^2} \approx z_o + \frac{x^2}{2z_o} + \frac{s^2}{8z_o} + \frac{sx}{2z_o} \quad (3.54)$$

Because the distance z_o is large compared to the other distances in the problem, we can expect the waves from the two slits to be essentially equally strong by the time they reach the screen, or at least we can use this as a simplifying approximation in our analysis here. For mathematical convenience, and also because quantum mechanical waves may well actually have this form, we use the complex exponential form of the solutions to the relevant wave equation here; that is, for a wavelength λ and hence a wavevector magnitude $k = 2\pi / \lambda$, we take the wave coming from a specific slit to be of the form $\exp(ikr)$ for some distance r from a source.

So the sum of the two waves from the two different slits, at the point x on the screen, will be approximately

$$\psi \propto \exp(ikd_1) + \exp(ikd_2) \quad (3.55)$$

Substituting from Eqs. (3.53) and (3.54) and writing

$$\alpha = k \left(z_o + \frac{x^2}{2z_o} + \frac{s^2}{8z_o} \right) \quad (3.56)$$

we will have

$$\begin{aligned}
 \psi(x) &\propto \exp\left[i\left(\alpha - k \frac{sx}{2z_o}\right)\right] + \exp\left[i\left(\alpha + k \frac{sx}{2z_o}\right)\right] \\
 &= \exp(i\alpha) \left[\exp\left(ik \frac{sx}{2z_o}\right) + \exp\left(-ik \frac{sx}{2z_o}\right) \right] \\
 &= 2 \exp(i\alpha) \cos\left(\frac{k sx}{2z_o}\right) = 2 \exp(i\alpha) \cos\left(\frac{\pi sx}{\lambda z_o}\right)
 \end{aligned} \tag{3.57}$$

Taking the “intensity” of the beam to be $\propto |\psi|^2$ (which would be essentially the probability density for a quantum mechanical wave), we would have a “brightness” on the screen of

$$|\psi_s(x)|^2 \propto \cos^2\left(\frac{\pi sx}{\lambda z_o}\right) = \frac{1}{2} \left[1 + \cos\left(\frac{2\pi sx}{\lambda z_o}\right) \right] \tag{3.58}$$

This corresponds to a pattern of interference “fringes” on the screen, with a period in the x direction of

$$d_s = \frac{\lambda z_o}{s} \tag{3.59}$$

We see, incidentally, that these interference fringes are not spaced by the wavelength, but by an amount that is z_o / s larger. Hence Young’s slits experiment allows us to measure wavelength without having any measuring device on the scale of the wavelength. The interference phenomena and the fringe spacing are shown in Fig. 3.12.

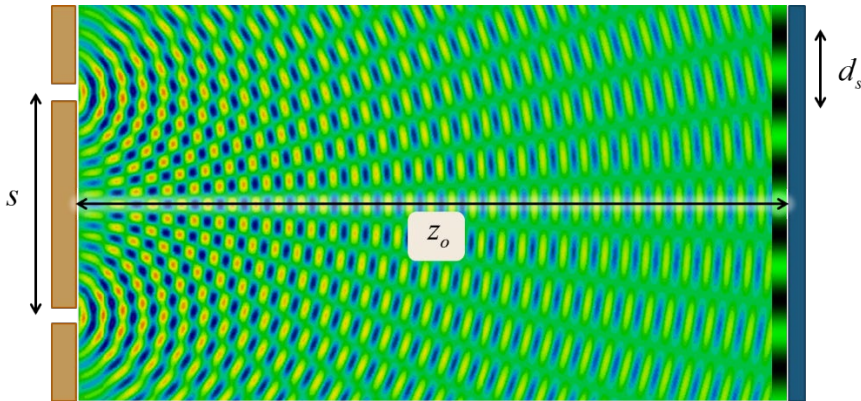


Fig. 3.12. Calculated interference pattern and resulting bright and dark “intensity” fringes spaced by a distance d_s on the screen on the right for Young’s slits when the slits are illuminated by a plane wave from the left.

Young’s slits and quantum mechanics 7c

Considered as a wave phenomenon, the behavior of Young’s slits is straightforward to understand. The apparent conceptual problems start when we want to think in terms of particles. Quantum mechanics is introducing for us the idea that entities such as the

photon and the electron have aspects of both wave and particle character – what we are referring to as “wave-particle duality”.

On the face of it, there might seem to be no problem if we are thinking of large numbers of photons, for example, because we might reasonably expect them to behave as waves, as we see every day in the normal “classical” behavior of radio waves and light beams from lasers. But what happens when we presume that we are dealing with very small numbers of photons or electrons? Specifically, what happens when we turn down the light beam intensity or the electron current so much that we can be quite sure that there is never more than one photon or electron in the apparatus?

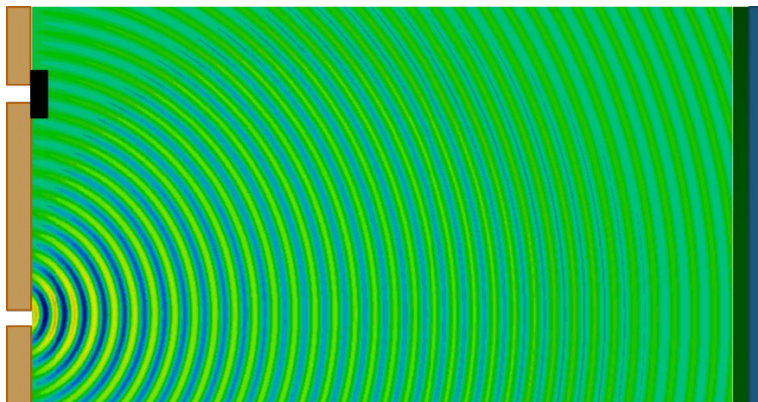


Fig. 3.13. Calculated diffraction pattern when the top slit is blocked.

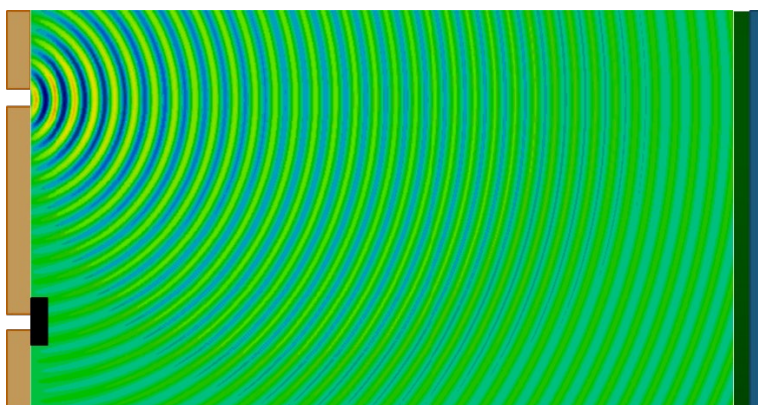


Fig. 3.14. Calculated diffraction pattern when the bottom slit is blocked.

We can readily calculate what happens even at high intensities or currents if we block one or other of the slits, as illustrated in Fig. 3.13 and Fig. 3.14. In either case, we would still see the wave diffracting as it emerges from the remaining aperture, but we would lose the interference “fringe” pattern on the screen.

Our classical intuition might therefore suggest to us that, when there is only one particle in the apparatus at a time, it will go through one or other of the slits, and hence we will lose the interference fringes. But, that is not what happens. In both the electron and the photon case, we still see the interference fringes. Now, at these low intensities or currents, for any one photon or electron, we will not see such a pattern. For a

photographic plate for a photon or for a phosphorescent screen for an electron, for any one photon or electron, we will just see one dot. The remarkable experimental conclusion³⁰, however, is that, if we keep repeating such two-slit experiments with one photon or electron at a time, the set of dots we get will progressively build up the interference pattern.

But how can this be? Surely, we might say, the electron or photon has to go through one slit or the other, so it cannot form an interference pattern. We remember, though, that when discussing the Bohr model of the atom, we made the point that the electron is not a “point particle”. Now we can extend that discussion, using this two-slit experiment to clarify how we view what is going on here. To do so, we need to introduce an idea that more typically only comes up in philosophical discussions but that will be helpful to us here.

The ontology of quantum mechanical particles

If we look up the word “ontology” in a dictionary, we will find a definition like “the nature of being”, which may not be very illuminating even though it is correct. More pragmatically, we can think of the ontology of something as being the set of attributes it possesses. A ghost can walk through walls. I cannot walk through walls. This distinction between me and a ghost is an ontological one. The ghost has the attribute of being able to walk through walls, and I do not have that attribute.

When we were making the transition between a classical view and a quantum mechanical one, we quite naturally used classical words to describe the new entities we were observing. So, for example, we called the electron a “particle”, and similarly for the photon. The mistake here is that we therefore also brought along all the ontological attributes of a classical particle when we did that, even though in fact we had no justification for doing so.

A classical particle has attributes like charge, mass, position, size, shape, and momentum. But just because we call an electron or a photon a particle does not mean that it has all of these attributes, and it may have others as well (spin is an example of such an additional attribute). With the benefit of hindsight, it might have been better to introduce another word to describe entities such as the electron and the photon; for example, maybe we should have referred to them as “qwarticles” or some other such new name. By doing so, we might have been able to avoid automatically bringing along all the classical attributes and hence causing ourselves a great deal of self-inflicted confusion.

How should we think about the attributes of these “qwarticles”? We can find experimentally that it is useful to give them the attributes of charge and mass (both of which are zero for the photon, but that “zero” value is still quite meaningful). We will need to add the attribute of spin, which the classical particles do not have. Charge, mass, and spin are indeed intrinsic attributes of electrons and photons.

³⁰ An excellent example of such an experiment performed with electrons can be seen at <http://www.hitachi.com/rd/portal/research/em/doubleslit.html>.

But we need to be much more careful about position, size, shape and momentum. We could ascribe some intertwined version of these attributes as allowed by the uncertainty principle. It is likely more useful and meaningful to say that the *state* that the “qwarticle” is in can have these properties in some sense. So, the hydrogen atom orbitals can have some kind of size and shape, at least in some average sense appropriate for a “fuzzy” object, and some concepts of position and momentum, consistent with the uncertainty principle. Those orbitals also possess quite specific angular momentum values. So we can say the electron is in such a state (a hydrogen atom orbital) that has these attributes; these attributes are really of the state, however, not of the idea of an electron itself.

This idea of distinguishing the ontology of the particle and the ontology of the state it is in is one we can use in the classical world as well; we can ascribe the position and momentum to the “state” of a classical particle, like a brick, for example, rather than having those be an intrinsic property of the brick itself. In the classical case, though, we would probably consider the shape to being an intrinsic property, whereas for the “qwarticle” the shape is better considered to be an attribute of the state the quarticle is in.

For the case of a photon, we can think of a “mode” of the electromagnetic field, such as some standing wave mode of a resonator, or the propagating “mode” of a laser beam. A photon can “occupy” that mode in the same way that an electron can occupy one of the eigenstates of the particle in a box or one of the orbitals of the hydrogen atom. As we have been discussing, these ideas of modes, eigenstates and orbitals are really all the same mathematical idea. We could simply use the word “mode” for all of them, though historically these different areas of physics have grown up somewhat independently so we do not typically use the word “mode” for all these applications in practice.

When we look at the diffraction pattern for Young’s slits, the pattern we calculate and plot is the “state” that is being occupied by the electron or the photon. It is not meaningful to ask where the electron or the photon is, or what slit it passed through³¹. These are not point particles that have to have a position. They are occupying a state that has a particular form and shape in space, and that size and shape encompasses both slits. Asking which slit the electron came through is asking for the value of an attribute that the electron does not intrinsically possess and for which the state it is in also obviously does not contain an answer. The question is meaningless. It is like asking someone to tell you where exactly a computer program is, to a precision of, say, 1 micron. The computer program is a state of some substantial part of the memory of your computer, likely much larger than 1 micron in size. Beyond that, it is meaningless to ask where it is.

The ontology of quantum mechanical waves

A slightly more subtle point, and one that is not normally discussed much in this context, is that the “wave” here does not have all the attributes of a classical wave either,

³¹ Incidentally, any attempt to figure out experimentally “what slit” the “particle” went through ends up by destroying the interference pattern.

and maybe we should actually refer to these also by another name we could make up, such as “qwave”³². Classical waves appear to be quite continuous objects, capable of any of a continuous range of amplitudes, for example, and those amplitudes are also real, measurable quantities. Our “qwaves” do have the attributes of propagation, linear superposition, and interference, just like many classical waves. But, quantum mechanical waves are not necessarily real, and they may not themselves be measurable entities. Furthermore, perhaps the attribute we might find most difficult to understand with quantum mechanical waves is how they can correspond to countable particles; classical waves do not appear to have any such attribute.

We have already seen with quantum mechanical waves that to get a direct quantitative meaning, such as probability density, we should “normalize” the wave, so that the probabilities add up to one. If we decide always to perform such normalization, then the wave becomes consistent in the sense that the particle always has to be somewhere when its position is measured³³. That is not an attribute we would require of a classical wave.

We have also seen that there may be an issue with quantum mechanical waves as to whether their amplitudes themselves are actually measurable quantities; furthermore, the amplitudes may not even be represented by real numbers. So for our quantum mechanical wave, it may not have the attribute that the wave can take any amplitude (it should perhaps always be normalized), and it may not have the attributes that it is even measurable or real. Even if we do not give the quantum mechanical wave (or “qwave”) such attributes, it is apparently still a useful entity as part of our quantum mechanical description of the world, however.

With the revised ontologies we are giving to our “qwarticles” and “qwaves”, we can now quite reasonably assert that there are no contradictions in “qwave-qwarticle” duality, and the ones we thought were there in “wave-particle” duality were because of our unjustified carrying over of the entire classical ontologies of waves and particles. While this is a quite reasonable and defensible attitude in quantum mechanics, and we recommend that the student takes this view, we have to be honest that we have also avoided discussing so far another area in which there are genuine difficulties in quantum mechanics, which is the issue of the measurement problem.

The measurement problem

The problem we have not dealt with so far lies in the observation that, in any such experiment with only one electron or one photon in the apparatus, the result of the

³² The reader should be clear that these terms “qwarticle” and “qwave” are just words we are making up here for the purposes of this argument. These are not words that are in use in quantum mechanics as far as the author is aware.

³³ Incidentally, how we handle quantum mechanical states for more than one particle is not that we simply increase the amplitude of the wave. In fact, to handle just two particles in three dimensional space, we need a wavefunction of 6 coordinates, three for each particle, and so on. Somewhat confusingly, that two-particle wavefunction would also be normalized so the probabilities add up to one, not two. The probabilities it describes, though, are the probability of finding the first particle at one position *and* the second particle at a second position.

experiment is to leave a dot on the screen, not an interference pattern (even though the pattern will build up if we keep repeating the experiment). Why does the experiment leave a dot at quite a definite position on the screen if we asserting that the electron or photon does not have a definite position inside the apparatus? Indeed, this is a perfectly fair question, and we would usually consider it strong evidence that the particle must have had a definite position when it hit the screen.

Essentially, we are asking for an explanation of Born's postulate that the modulus squared of the wavefunction gives us the probability of finding the particle at a point. We could "explain" this by saying it is because of the "collapse of the wavefunction". That is about as good an explanation as saying that something is hot because it has a high temperature – it categorizes the problem to some degree in scientific terms, but offers no explanation of any process.

The idea of the "collapse of the wavefunction" is saying that, though the wavefunction was distributed through some region of space before the measurement, the act of measuring the position causes the wavefunction to "collapse" to being concentrated about some point. A more generalized statement of the same idea is to say that

the act of measuring some quantity causes the system to collapse into an eigenstate of the quantity being measured

with probabilities given by Born's rule or a generalized version of it³⁴. Here, since we are measuring position, we note that we are "collapsing the wavefunction" into one with a definite position.

To be clear, we are not explaining Born's rule or the supposed mechanism of how this collapse takes place. The fact that we are not at all clear on how to do either of these things is what is known as the "measurement problem".

This measurement problem is quite a deep one in quantum mechanics. We can actually prove that a simple application of quantum mechanical rules does not allow us to describe any machine that can cause the collapse of the wavefunction; in other words, not only are we saying that we cannot describe the details of this process, we can actually prove that, according to quantum mechanics as we usually state it, the collapse of the wavefunction cannot actually happen!

Of course, this collapse does happen, so we are left with a problem here. Not surprisingly, there are various proposed approaches to getting us out of this difficulty. A short list of the names of these various approaches includes the Copenhagen interpretation, Bohm's pilot wave, nonlinearity in quantum mechanics, and the "many-worlds" hypothesis. Though each of these approaches (and various others also) have their advocates, it would be accurate to say that there is no one approach that is universally accepted and that none of these is easy to explain convincingly.

As a practical matter, what essentially everyone has to do is merely accept Born's rule as being something that works for all actual calculations and prediction, even if this leaves us in the unsatisfactory condition of not really understanding why it works. The

³⁴ The generalized statement would be "with probabilities given by the modulus squared of the expansion coefficients of the state on the eigenfunctions of the quantity being measured".

good news is that, apparently, it works very well if we merely presume that measurement is something we do with large apparatuses on small objects. Indeed, if we had an experiment that could show that this postulate did not work in some set of circumstances, we could perhaps figure out how to resolve this difficulty. In the meantime, we may just have to give this practical approach its technical name, widely accepted across the physics community, which is “shut up and calculate”.

Quiz questions

- Q3.6.1 Consider a slit or aperture of width d in an otherwise opaque screen. We are shining light of a specific wavelength λ on the “back” of the screen and are looking at the diffracting light that emerges from the aperture, and are interested in the diffraction angle θ_c of that light.
- a) If I increase the wavelength λ , state whether the diffraction angle θ_c increases, decreases, or stays the same.
 - b) If I increase the aperture width d , state whether the diffraction angle θ_c increases, decreases, or stays the same.
 - c) If I increase the magnitude of wavevector k , state whether the diffraction angle θ_c increases, decreases, or stays the same.
 - d) If the light has wavelength $\lambda = 1$ micrometer and the slit or aperture has a width of $d = 3$ micrometers, which one of the following would be a reasonable estimate of the characteristic half-angle of the diffracting light emerging from the slit or aperture? 3° , 27° , or $(1/1080)^\circ$? (Note: Unless otherwise stated, angles in formulas are generally in radians, not degrees. To convert from radians to degrees, multiply by $360/2\pi$.)
 - e) If I increase the magnitude of the photon momentum p of the photons in the incident light beam, state whether the diffraction angle θ_c increases, decreases, or stays the same.
 - f) Now we explicitly consider the light initially to be a plane wave propagating in the z direction, and we consider the width of the slit or aperture to be a width in the x direction (perpendicular to the z direction). For the light that passes through the slit or aperture, we are interested in the uncertainty Δp_x in the magnitude p_x of the component of the photon momentum in this x direction. If I increase the width d of the slit, state whether the uncertainty Δp_x increases, decreases, or stays the same.
- Q3.6.2 If Young’s slits have a separation of 100 microns, the distance from the slits to the screen is 30 cm, and the separation between the centers of adjacent bright fringes on the screen is 15 microns, what is the wavelength of the light?
- Q3.6.3 For an electron, for each of the following attributes, choose whether, in quantum mechanics, we assign the attribute to being an attribute of the state or to being an intrinsic attribute of the electron (i.e., an attribute whose value is the same for all electrons).

- a) Mass
- b) Momentum
- c) Charge
- d) Shape
- e) Location (i.e., where we are likely to find the particle)

Q3.6.4 For a photon, for each of the following attributes, choose whether, in quantum mechanics, we assign the attribute to being an attribute of the state or to being an intrinsic attribute of the photon (i.e., an attribute whose value is the same for all photons).

- a) Energy
- b) Charge
- c) Wavelength
- d) Momentum
- e) Mass

Problems

3.6.1* Suppose we have a particle whose position we can know within an uncertainty of $\pm\Delta x$ where $\Delta x = 10 \text{ nm}$. We presume the particle has a mass m . If it has an uncertainty in its momentum Δp_x in the x direction, we will presume that corresponds to an uncertainty in its velocity in the x direction of $\Delta v_x = \Delta p_x / m$. Presuming these quantities Δx and Δp_x are appropriately defined to correspond to the usual statement of Heisenberg's uncertainty principle, what is the minimum uncertainty Δv_x we should expect in the velocity in the x direction of a such a particle in each of the following cases? (Give answers to three significant figures, in m/s.)

- (i) If the particle is an electron.
- (ii) If the particle is a proton.
- (iii) If the particle is a cube of side 1 micron in length of a material of density 10 g/cm^3 .

3.6.2 A neutron moving in x direction is passing through a slit of width $10 \text{ }\mu\text{m}$ in the y direction. Which of the following is the closest approximate estimate of the minimum uncertainty of its momentum in the y -direction given that it passes through the slit?

- a) $\Delta p_y \approx 10^{-43} \text{ kg m/s}$,
- b) $\Delta p_y \approx 10^{-29} \text{ kg m/s}$,
- c) $\Delta p_y \approx 10^{-14} \text{ kg m/s}$,
- d) $\Delta p_y \approx 1 \text{ kg m/s}$.

3.6.3* Consider a two-slit interference experiment, as in Young's slits. So we have a mask, oriented parallel to the x axis, that is opaque except for two parallel slits.

Presume the slits are very narrow, and are separated in the x direction by a distance of 10 microns. A screen, parallel to the mask, is positioned away from the mask in the z direction, separated from the mask by a distance $z_o = 1 \text{ m}$. We expect in an experiment where the slits are illuminated by a plane wave traveling along the z direction to the back of the mask that we will see interference fringes on the screen. What will be the spacing between the centers of adjacent dark fringes on the screen in the following situations?

- (i) In an optical version of the experiment, where the incident beam corresponds to photons with an energy of 1 eV.
- (ii) In a version of the experiment with monoenergetic (i.e., monochromatic) electrons with a kinetic energy of 1 eV.

3.6.4 Consider a “Young’s slits” experiment for an electron. We are going to generate an appropriate electron wave for our experiment by first starting off with some electrons that we somehow manage to generate with essentially zero kinetic energy (for example by managing to emit them from a metal “cathode” by the photoelectric effect at a carefully chosen optical wavelength, though the details of this do not matter for our question). Then we are going to accelerate them towards a “mask” (which would also be the “anode” in this experiment) using an appropriate electrostatic potential (voltage) V_a . (For example, the metal from which we emit the electrons might be at 0V, and the mask itself might be a metal held at a voltage V_a .) The mask has two very narrow, parallel slits cut in it, separated by 10 μm , and is otherwise opaque to the electrons (i.e., electrons can only pass through the slits). (The whole apparatus here is in a vacuum.)

We want to form an interference pattern on a screen, parallel to the mask plane, and at a distance 1 m from the mask (on the other side from the electron source). What voltage V_a should we apply if the separation between the “bright” fringe peaks on the screen is to be 0.5 mm?

3.7 Tunneling

In the classical world, we can imagine a ball that we start off with some kinetic energy, and that then encounters a hill that slopes upwards. We know that, as the ball starts to roll up the hill, its speed gradually reduces as some of its kinetic energy gets turned into potential energy. The ball will only be able to get to the other side of the hill if its initial kinetic energy exceeds the potential energy it would have at the top of the hill³⁵.

In quantum mechanics, however, it is possible for a particle to get to the other side of a barrier that is “too high” in this sense. Even if the energy of the particle is lower than the height of the barrier, it is possible for the particle to get to the other side. This is a process that happens with some probability, and the higher and thicker the barrier is, the lower is the probability.

³⁵ Or, at least, that of the highest point it will encounter on its path up and over the hill.

This process by which the quantum mechanical particle gets to the other side of a barrier that is nominally too high for it is called tunneling – loosely, it is as if the particle has somehow managed to tunnel through the hill – though it is important to emphasize that this process is not at all like the classical idea of digging a tunnel.

This process of tunneling is quite common in the quantum mechanical world, and is found routinely in modern electronic devices. In so-called “flash” memories that work by storing charge semi-permanently on small capacitors, the charge is put on and taken off the relevant capacitor plate by “tunneling” electrons through nominally insulating dielectrics. In modern MOS (metal-oxide-semiconductor) transistors, the tunneling of electrons through the nominally-insulating gate dielectric is an undesired nuisance. Gate dielectrics have become progressively thinner as transistors have become progressively smaller. Tunneling of electrons through those gate dielectrics has become a substantial problem. This has led to a major change to different gate dielectric materials, such as hafnium dioxide, as an alternative to the previous silicon dioxide. These allow the use of thicker gate dielectrics, thereby reducing the tunneling, while retaining the necessary capacitance between the gate and the channel that allows the transistor to be turned on and off by changing the voltage on the gate.

Though tunneling might seem a very strange process, it is actually quite straightforward to describe it using quantum mechanical waves and Schrödinger’s wave equation. As we have seen before, quantum mechanical waves show many of the behaviors we see for classical waves. The process of tunneling also exists for classical waves, with the best known examples being in what are called “evanescent fields” in optics.

Such evanescent fields are seen near to a surface when an incident wave arrives at an angle past the so-called critical angle for total internal reflection. We can see such total internal reflection as we try to pass from a material of high refractive index to one of low refractive index. If you can swim underwater with your eyes open, you are likely quite used to the fact that there is only a moderate range of angles at which you can see “through” the surface of the water to the world above. Outside that range of angles, the surface of the water appears to be totally reflecting, like a mirror. If you were to shine a flashlight from below the surface of the water at an angle such that it was totally reflected in this way back into the water, there would be an evanescent field just above the surface of the water, a field that decays very quickly with distance. If, however, you were to put a piece of glass just above the surface of the water, but not touching it, within the short range of these evanescent fields, the light from the flashlight would be able to “tunnel” into the glass, forming a light beam there in the glass above the water surface, even though the glass was not touching the water surface. This phenomenon – known in optics as “frustrated total internal reflection” – mathematically corresponds to the light “tunneling” through the air gap between the water and the glass, and it is used to make some kinds of partially reflecting mirrors, often known as “beam splitters”, with two closely spaced pieces of glass, for example.

To understand tunneling in quantum mechanics, we can start by looking at what happens when a quantum mechanical wave arrives at a potential barrier of finite height.

A potential barrier of finite height

A potential barrier of finite height is sketched in Fig. 3.15. For simplicity, this barrier is at the position $z = 0$, and the barrier corresponds to a potential energy V_o relative to the potential energy for $z < 0$, which we choose to be 0.

Nature of solutions for a finite barrier

Quite generally, in the region $z < 0$, we can write the solution to Schrödinger's time-independent wave equation in the form

$$\psi_{\text{left}}(z) = A \exp(ikz) + B \exp(-ikz) \quad (3.60)$$

where A and B are arbitrary constants (which may be complex numbers) that might be fixed later once we apply some boundary conditions or other constraints to the problem of interest.

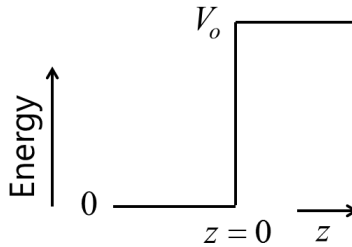


Fig. 3.15. A potential barrier of finite height, V_o , at position $z = 0$. The positive z direction runs to the right.

In quantum mechanics, we interpret a solution of the form $\exp(ikz)$ to correspond to a wave propagating in the positive z direction, and, similarly, a wave of the form $\exp(-ikz)$ corresponds to one propagating in the negative z direction³⁶. So, in Eq. (3.60) $A \exp(ikz)$ corresponds to a wave propagating to the right with amplitude A , and $B \exp(-ikz)$ is a wave propagating to the left with amplitude B .

³⁶ This choice comes from a choice made in considering time-dependent behavior, and specifically, Schrödinger's time-dependent wave equation, which leads to the time behavior being written in the form $\exp(-i\omega t)$ with positive ω . The reader might also reasonably ask why we write the solution here in this form $\psi_{\text{left}}(z) = A \exp(ikz) + B \exp(-ikz)$ rather than the form we had in Eq. (3.33) – a form like $\psi_{\text{left}}(z) = F \cos(kz) + G \sin(kz)$. Both of these forms are general solutions, and both are correct. Because of Euler's formula $\exp(i\theta) = \cos \theta + i \sin \theta$, $F \equiv A + B$, and $G \equiv i(A - B)$, and both forms are equivalent. The cosine and sine form, however, is describing the wave in terms of a sum of two standing waves, whereas the complex exponential form is describing the wave in terms of the sum of two “running” waves, one in each direction. It may be surprising mathematically that both forms are capable of describing either running or standing waves or any combination of the two, but, because of the time behavior is a multiplicative factor $\exp(-i\omega t)$, this turns out to be correct. For a forward propagating wave $\exp[-i(\omega t - kz)] \equiv \exp(-i\omega t) \exp(ikz) \equiv \exp(-i\omega t)(\cos kz + i \sin kz)$. The “running wave” form is more convenient for us here, so we use it for the moment.

The barrier on the right gives boundary conditions that restrict the relation between A and B . In this problem, we do not presume any barrier on the left. As a result, there is no constraint on the allowed energy E , at least if we choose it to be positive. So, we simply have

$$k = \sqrt{\frac{2mE}{\hbar^2}} \quad (3.61)$$

Now we need to consider the effect of this finite barrier on the solutions. To understand this, we look first at the nature of the solutions to Schrödinger's wave equation for $z > 0$, that is, inside the barrier, on the right in Fig. 3.15. Mathematically, in that region, Schrödinger's equation becomes

$$-\frac{\hbar^2}{2m} \frac{d^2\psi(z)}{dz^2} + V_o\psi(z) = E\psi(z) \quad (3.62)$$

which we can rewrite as

$$\frac{d^2\psi(z)}{dz^2} = \frac{2m}{\hbar^2} (V_o - E)\psi(z) \quad (3.63)$$

Suppose now we presume that the energy associated with the state – that is, E – is below the height of the top of the barrier (i.e., $E < V_o$), so the quantity $V_o - E > 0$. We can define a real quantity κ (the Greek letter “kappa”)

$$\kappa = \sqrt{\frac{2m(V_o - E)}{\hbar^2}} \quad (3.64)$$

which we can choose to be positive, and which has dimensions of inverse distance.

With these choices, we can write, for this region inside the barrier, to the right,

$$\frac{d^2\psi(z)}{dz^2} = \kappa^2\psi(z) \quad (3.65)$$

The general solution to this equation is of the form

$$\psi_{right}(z) = C \exp(-\kappa z) + D \exp(\kappa z) \quad (3.66)$$

For the situation as in Fig. 3.15, where we presume that the barrier goes on forever on the right – that is, it continues for arbitrarily large z – we can decide that we should choose $D = 0$; otherwise the solution will grow arbitrarily large with increasing z , which we could regard as being unphysical³⁷. Hence we presume we can have a solution of the form

$$\psi_{right}(z) = C \exp(-\kappa z) \quad (3.67)$$

inside the barrier.

³⁷ If we consider a situation with a barrier of finite thickness and presume there is no wave incident from the right, in the limit of a thick barrier, we can also justify presuming D becomes arbitrarily small.

Perhaps the first point we should note here is that there is a solution corresponding to a finite wave inside the barrier even for energy below the top of the barrier. This is already different from what we might expect classically for a particle³⁸. The second point we can note is that this solution corresponds not to a “propagating” wave, but to a decaying exponential.

Now that we have established the nature of the solutions on the left and on the right, we need to find out how to join them up. This joining requires boundary conditions.

Boundary conditions

We could regard abrupt boundaries like the one we have proposed here as being unphysical in that nothing happens so abruptly in the real physical world, and that would be a reasonable criticism. If, however, we work with potentials without such abruptness, the mathematics becomes harder. What we can do is to think of such abrupt potentials as being a mathematically convenient limit of some smoother potential and find some way to deal with the abrupt interface, which is where “boundary conditions” come in.

If we consider the Schrödinger equation quite generally, as in

$$-\frac{\hbar^2}{2m} \frac{d^2\psi(z)}{dz^2} + V(z)\psi(z) = E\psi(z) \quad (3.68)$$

we can presume that, in any real problem we want E and V always to be finite, and we might similarly presume that we want ψ to be finite also. In that case, Eq. (3.68) tells us that $d^2\psi/dz^2$ must also be finite, because all the other terms in the equation are finite by presumption. Since the second derivative is the rate of change of the first derivative, we cannot have any discontinuous jumps in the first derivative since otherwise at that point $d^2\psi/dz^2$ would be infinite. So we require the boundary condition

$$\frac{d\psi(z)}{dz} \text{ is continuous} \quad (3.69)$$

Also, $d\psi/dz$ must be finite too, since otherwise $d^2\psi/dz^2$ could be infinite, being the limit of a difference involving infinite quantities. For $d\psi/dz$ to be finite, then we similarly require

$$\psi(z) \text{ is continuous} \quad (3.70)$$

We could regard the above arguments on these boundary conditions as being informal and not entirely rigorous; we could again justify them more rigorously by considering them as being the limit of what happens with smooth potentials as they tend towards abrupt ones. That would take some time and mathematical care, however, so here we merely presume we can use these boundary conditions, Eqs. (3.69) and (3.70), which are anyway quite common ones in differential equations.

³⁸ We do, however, also see such penetration effects with classical waves, such as the evanescent optical waves discussed above.

Solutions for a barrier of finite height

Using these boundary conditions, we have just proposed, we can now attempt to join the solutions, Eqs. (3.60) and (3.67), in the two regions. First, we can use the continuity of the wavefunction – that is, condition (3.70) – at $z = 0$ to give us, from Eqs. (3.60) and (3.67),

$$\psi_{\text{left}}(0) = A + B = \psi_{\text{right}}(0) = C \quad (3.71)$$

The derivatives of the wavefunction forms are straightforward to evaluate, giving

$$\frac{d\psi_{\text{left}}(z)}{dz} = \frac{d}{dz} [A \exp(ikz) + B \exp(-ikz)] = ikA \exp(ikz) - ikB \exp(-ikz) \quad (3.72)$$

and

$$\frac{d\psi_{\text{right}}(z)}{dz} = \frac{d}{dz} C \exp(-\kappa z) = -\kappa C \exp(-\kappa z) \quad (3.73)$$

So, at $z = 0$, equating the derivatives on the left and on the right of the barrier,

$$ikA - ikB = -\kappa C \quad (3.74)$$

or equivalently

$$A - B = \frac{-\kappa}{ik} C = \frac{i\kappa}{k} C \quad (3.75)$$

Adding Eqs. (3.71) and (3.75) gives

$$2A = \left(1 + \frac{i\kappa}{k}\right) C = \left(\frac{k + i\kappa}{k}\right) C \quad (3.76)$$

or equivalently

$$C = \frac{2k}{k + i\kappa} A = \frac{2k(k - i\kappa)}{k^2 + \kappa^2} A \quad (3.77)$$

Subtracting these same equations similarly allows us to establish a relation between B and C , and hence, using Eq. (3.77), a relation between A and B . Establishing both of these relations is left as a pair of exercises for the reader.

The key result we can see from Eq. (3.77) is that, for any finite forward wave incident on the barrier from the left – that is, for any finite value of A – there is a wave inside the barrier, in the form of a decaying exponential. We could think of this wave as “tunneling penetration” into the barrier.

We could plot up the results for the wavefunctions throughout this structure, using Eq. (3.60) for the waves to the left of the barrier, that is, for $z < 0$, and using Eq. (3.67) for the waves to the right of the barrier, where we would use Eq. (3.77) to relate the coefficients A and C . A similar equation would relate B and C . So, within one arbitrary constant (which we could choose to be either A or C) we would know the entire wavefunction in this problem. Here we do not attempt to normalize the wavefunction,

which would set the remaining arbitrary constant. This is in part because such normalization would depend on what happens on the far left, and we are not for the moment interested in that region.

Such a wavefunction would be a complex entity. It is easier and more meaningful to plot the probability density (at least in arbitrary units because we have not normalized), which is a real number. To find the probability density corresponding to these solutions, in each region we have to take the modulus squared of the entire wavefunctions. That is, on the left

$$|\psi_{\text{left}}(z)|^2 = |A \exp(ikz) + B \exp(-ikz)|^2 \quad (3.78)$$

and on the right

$$|\psi_{\text{right}}(z)|^2 = |C \exp(-\kappa z)|^2 = |C|^2 \exp(-2\kappa z) \quad (3.79)$$

An example calculation of the probability density is shown in Fig. 3.16.

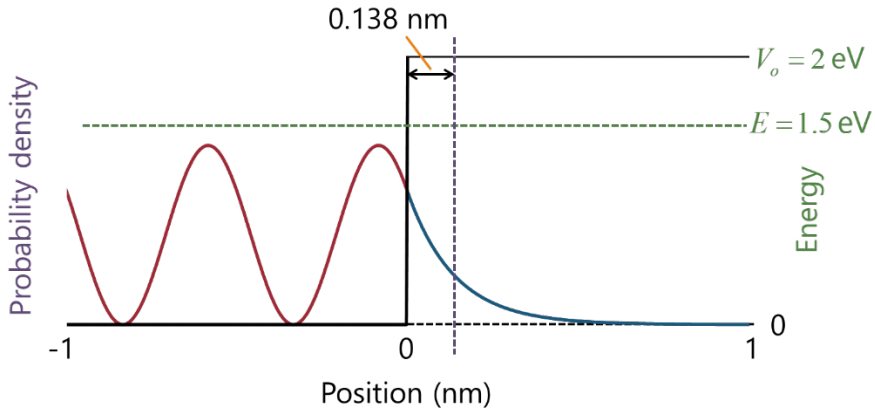


Fig. 3.16. Calculated probability density for the potential structure shown. The potential barrier is of height $V_o = 2 \text{ eV}$ and an electron incident from the left has energy $E = 1.5 \text{ eV}$. The calculated magnitude of the wavevector on the left is $k \approx 6.275 \times 10^9 \text{ m}^{-1} \equiv 6.275 \text{ nm}^{-1}$, which corresponds to a wavelength of $\lambda \approx 1.001 \text{ nm}$, so half a wavelength, which corresponds to the separation of the peaks or of the minima in the standing wave on the left, is $\approx 0.50 \text{ nm}$. The exponential decay constant for the wave on the right is $\kappa \approx 3.623 \times 10^9 \text{ m}^{-1} \equiv 3.623 \text{ nm}^{-1}$, so the $1/e$ decay length of the probability density on the right is $1/2\kappa \approx 0.138 \text{ nm}$.

Note that we see the exponential decay of the probability density into the barrier on the right. We also see a “standing wave” on the left. An infinitely thick barrier such as this has 100% reflection for incident energies lower than the height of the barrier, so we see the incident wave is completely reflected even though there is some penetration into the barrier. This is just like the situation of shining a flashlight inside the water up to the surface of the water; beyond the critical angle for total internal reflection, all the light is reflected, even though the light penetrates a very small distance, in a similar exponential fashion, into the air above. Because the reflection is 100% in magnitude,

the standing waves we see on the left have minima of zero because the forward and backward waves, being of equal magnitude, can exactly cancel one another there by interference. The separation between the minima is half the wavelength on the left. Note also that the phase change on reflection is neither 180° , which would lead to a zero in the wave at a reflecting boundary, nor 0° , which would lead to a maximum at a reflecting boundary. The phase change on reflection here is some other specific value; such a different mathematical value is required so that we match both the amplitude of the wave and its derivative as we join up with the decaying exponential wave in the barrier.

A barrier of finite height and thickness 9c

If we make a barrier of finite thickness L_b , then we can see tunneling through the barrier. To analyze such a case, we need to consider three regions – a region on the left with forward and backward propagating waves, a region inside the barrier, where now we allow for both decaying and growing exponentials, and a region to the right where we presume a forward propagating wave only.

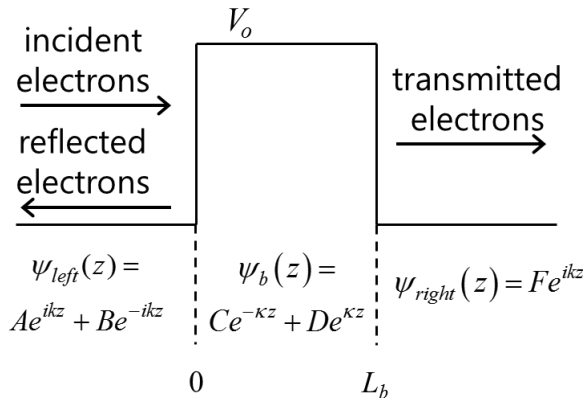


Fig. 3.17. A finite barrier structure and the forms of the waves inside each region when the incident electron energy E is below the height V_o of the barrier.

Note that because this barrier is of finite thickness, the argument we used before for discarding the exponentially “growing” term no longer applies; there is now no reason to throw this term away. In fact, this exponentially growing term can be regarded as a “backwards” exponentially decaying term that results from reflection of the “wave” inside the structure from the right side of the barrier.

One way to analyze such a structure is to work backwards from the right. We could choose any value we like for the amplitude F of the forward propagating wave on the right for the moment, and work out everything else in terms of it.

We can apply the boundary conditions on the right interface to establish relations between the amplitudes C and D of the “forward” and “backward” exponentially decaying waves inside the barrier and the chosen F . Then at the left interface, we can use the boundary conditions to establish the relations between the forward and backward propagating wave amplitudes, A and B respectively, on the left and the

amplitudes C and D . In this way we can relate all of the constants A , B , C , and D to the arbitrarily chosen F . Again, we are not normalizing the wave here because we have not chosen what happens on the far left or the far right, and we are not for the moment interested in those regions.

Note also, because we are interested in tunneling from the left to the right, we presume that there is only a forward-propagating wave on the right; we presume there is no electron wave incident from the right.

The probability density in the incident forward wave is proportional to $|A|^2$. The probability density in the “transmitted” forward wave is proportional to $|F|^2$. The situations are otherwise similar on the left and the right – for example, the potential energy is the same. Consequently, we can think of the fraction of the incident “current” that is transmitted by the barrier, or equivalently, the probability that a particle incident with this energy E on the barrier will be transmitted through the barrier, as $|F|^2 / |A|^2$.

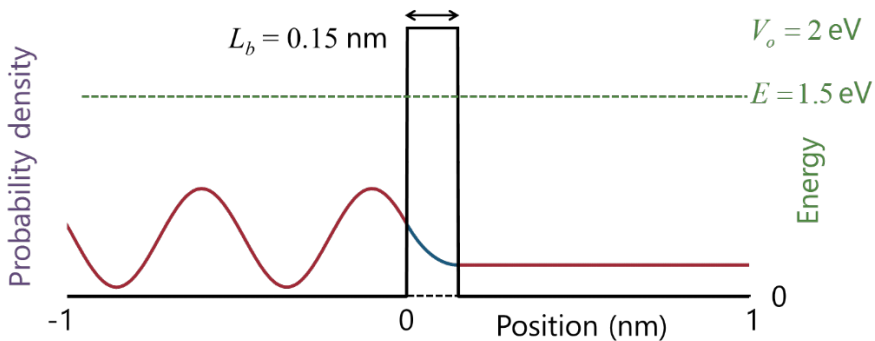


Fig. 3.18. Calculated probability density for the potential structure shown, with an electron wave incident from the left. The parameters and scales are the same as in Fig. 3.16 above except that the barrier is $L_b = 0.15$ nm thick.

The algebra for solving this particular problem is left as an exercise for the reader. In Fig. 3.18, we show the results of such a calculation. The parameters here are the same as in Fig. 3.16, except we have chosen the barrier to have a thickness $L_b = 0.15$ nm.

Note several aspects of the resulting behavior in Fig. 3.18. First, we see now that there is a finite probability density on the right – the wave is indeed tunneling through this barrier. The behavior inside the barrier is still predominantly an exponential decay, though now the backwards exponential decay is not negligible, so the shape is actually the sum of forward and backward exponential decays (with a smaller amplitude on the backward one).

On the left, we are still seeing “standing wave” behavior in the probability density, with the same period. Note now that the minima in the standing wave do not go down to zero. This is because the reflection off the barrier is now not 100%; there is partial transmission through the barrier. So the backward propagating wave on the left has a smaller amplitude than the forward propagating wave. Hence the interference between these two waves does not lead to complete cancellation at the minima. The standing wave maxima in Fig. 3.18 are also lower than those in Fig. 3.16 for the same reason;

Now the sum of the amplitudes of the forward and backward propagating waves is not as large at its peaks. Note also that since the probability density is the modulus squared of the total wave, even a small drop in the amplitude of the backward propagating wave leads to a significant drop in the maxima of the probability density.

Reflected waves, particles, and energy

There is one point that often causes confusion in thinking about quantum mechanical waves and reflections off barriers or other structures with potentials that vary in space; this point is one area where classical and quantum mechanical waves differ profoundly. That point has to do with how we think about particles and energy when quantum mechanical waves are reflecting or transmitting.

For classical waves, if we have some partial reflection off a boundary, such as light waves reflecting off glass, we think of some amount of transmitted power and some amount of reflected power. We can express that as a “splitting” of the power; for example, for glass perhaps 96% of the light is transmitted into the glass and 4% is reflected. That power splitting is also related to the amplitudes of the transmitted and reflected light waves in ways we can describe quite explicitly in electromagnetism, for example³⁹.

In quantum mechanics, as we have seen, the quantum mechanical wave associated with some particle, such as an electron, may similarly be partly reflected and partly transmitted. The key point, though, is that does not mean that the particle is “split” into a part that is reflected and a part that is transmitted; and it does not mean that the energy of the particle is “split” between the reflected and transmitted waves. Quantum mechanical wave amplitudes allow us to calculate relative probabilities of what happens to the particle; those probabilities are based on the modulus squared of the wavefunction. They can tell us the probability that the particle is transmitted or reflected. But, regardless of what those probabilities are, if we find the particle is transmitted, then we find the whole particle is transmitted, and we get the whole energy of the particle when we measure the energy of that transmitted particle. The same behavior applies for reflection; if we find the particle is reflected, the whole particle is reflected, and we get the whole energy of the particle.

It is still meaningful to talk about reflection and transmission coefficients for barriers; if we repeated the experiment with many particles, and measured whether the particle was reflected or transmitted, then on the average the percentages would correspond to those reflection and transmission coefficients.

With light beams, with large numbers of photons in the beams, we would find that the numbers of photons reflected and transmitted would correspond also with the classical calculation of reflection and transmission coefficients based on classical electromagnetism. The point there is that, if we sent just one photon into the apparatus, it would either be completely reflected or completely transmitted; the statistical average fraction reflected or transmitted is what would correspond to the classical calculation.

³⁹ Fresnel’s equations would tell us the reflection and transmission of the fields, and the Poynting vector would allow us to calculate the power in the reflected and transmitted beams.

Quiz questions

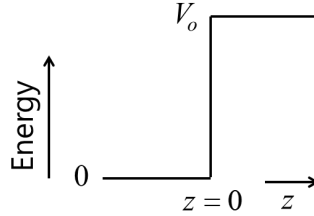
- Q3.7.1 I have a barrier with a given "height" V_0 in energy and of infinite thickness. The potential energy on the left side of the barrier can be taken to be uniform and zero. A particle (e.g., an electron) with energy $E < V_0$ is incident from the "left" side of the barrier.
- As I increase the energy E (but always with $E < V_0$, the probability of finding the particle in some specific region inside the barrier increases. True or False?
 - As I increase the energy (but always with $E < V_0$, the exponential decay length of the probability distribution inside the barrier increases.
 - As I increase the energy E (but always with $E < V_0$, though there is a standing wave pattern on the "left" of the barrier, it will no longer have points at which the probability density reaches zero.
 - As I decrease the energy towards $E = 0$, the probability of finding the particle inside the barrier also decreases asymptotically to zero.
- Q3.7.2 I have a barrier of a given "height" V_0 in energy and some finite thickness L_b . The potential energy on both sides of the barrier can be taken to be uniform and zero. A particle with energy $E < V_0$ is incident from the "left" side of the barrier. (No particle is incident from the right of the barrier.) State whether each of the following statements is true or false as I increase the width of the barrier.
- The probability that the particle is reflected increases.
 - There will be points on the "left" of the barrier where I am increasingly *less* likely to find the particle.
 - There will be points on the "left" of the barrier where I am increasingly *more* likely to find the particle.
 - There will be points on the "right" of the barrier where I am increasingly *less* likely to find the particle.
 - There will be points on the "right" of the barrier where I am increasingly *more* likely to find the particle.

Problems

- 3.7.1 An electron of kinetic energy in the z direction of 1 eV is incident on a potential barrier of arbitrarily large thickness and of height 1.2 eV.
- What is the exponential decay length of the electron's probability density (note, *not* of the quantum mechanical wave amplitude) inside the barrier? (That is, what is the distance over which the probability density decays to $1/e$ of its initial value, where e is the base of the natural logarithms?)
 - Compare the probability P_1 of finding the electron within the first 1 nm of depth into the barrier with the probability P_2 of finding the electron within

the second 1 nm of depth (i.e., between 1 nm and 2 nm into the barrier).
What is the ratio P_2 / P_1 ?

- 3.7.2 Consider a potential barrier of arbitrary thickness and of height V_o at a position $z = 0$ as in the figure below.



We presume an electron with kinetic energy E less than the height of the barrier corresponds to a wave on the left of the barrier of the form

$$\psi_{left}(z) = A \exp(ikz) + B \exp(-ikz)$$

and inside the barrier is of the form

$$\psi_{right}(z) = C \exp(-\kappa z)$$

where A , B , and C are constants that may be complex.

- (i) Derive an expression for the ratio $R = B / A$ in terms of k and κ , arranging that only the numerator is a complex number (that is, the denominator should be real).
 - (ii) If we write $R = |R| \exp(i\theta)$, derive an expression in terms of k and κ for $\tan \theta$.
 - (iii) Rewrite your expression for $\tan \theta$ in terms of E , V_o and fundamental constants.
- 3.7.3* Consider a potential barrier of arbitrary thickness and of height V_o at a position $z = 0$. Suppose now we have an electron incident on the barrier from the left, with energy E that is *greater* than V_o . Give an expression for the wave in the region $z > 0$ in terms of E , V_o and fundamental constants (though you may leave the overall amplitude of the wave arbitrary within a multiplicative constant). [Note: there is no wave coming from the “right” in this problem, so you can neglect any such wave in your solution.]
- 3.7.4 Consider a potential barrier of arbitrary thickness and of height V_o at a position $z = 0$, as in the preceding two questions, and presume an electron with kinetic energy $E > 0$ and less than the height V_o of the barrier. We have a wavefunction on the “left” of the barrier (i.e., for $z < 0$) of the usual form

$$\psi_{left}(z) = A \exp(ikz) + B \exp(-ikz)$$

where k has an appropriate value. Suppose that we have proved that

$$R = \frac{B}{A} = \exp(-i\theta)$$

where θ is a real number that we can calculate based on the energy E and the barrier height V_o . (Note that this means we have concluded, correctly, that the magnitude of this reflection is unity.) θ is then a measure of the “phase change on reflection” by the wave. Specifically, different values of θ would lead to different positions of the first maximum in the standing wave pattern for the probability density to the left of the barrier.

We are interested in the position z_1 of the first such maximum in the standing wave pattern for the probability density on the *left* of the barrier. (Note that z_1 will be less than zero because it is on the left of the barrier.) We can presume that $0 < \theta < \pi$.

Find an expression for z_1 in terms of E , θ , and fundamental constants.

- 3.7.5* Consider a potential barrier of finite thickness, L_b , and of height V_o at a position $z = 0$ as in Fig. 3.17. We presume an electron with kinetic energy E less than the height V_o of the barrier corresponds to a wave (as also shown in Fig. 3.17), on the left of the barrier of the form

$$\psi_{\text{left}}(z) = A \exp(ikz) + B \exp(-ikz)$$

inside the barrier of the form

$$\psi_{\text{left}}(z) = C \exp(-\kappa z) + D \exp(\kappa z)$$

and to the right of the barrier of the form

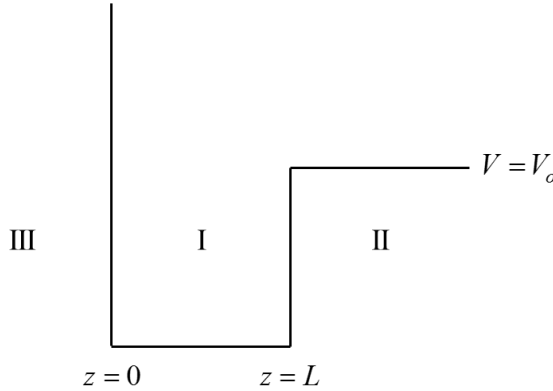
$$\psi_{\text{right}}(z) = F \exp[ik(z - L_b)]$$

where A , B , C , D , and F are constants that may be complex. Note we presume no backwards propagating wave on the right of the barrier. (Note the choice of the form $F \exp[ik(z - L_b)]$ on the right, which may make your algebra slightly easier. We are always free to make such a choice because F is an arbitrary constant to be determined, and $\exp(-ikL_b)$ is also a constant. Equivalently, $F \exp[ik(z - L_b)] \equiv G \exp(ikz)$ for some other arbitrary constant $G = F \exp(-ikL_b)$.)

- (i) Write down expressions for k and κ in terms of E , V_o and fundamental constants.
- (ii) Derive an expression for the ratio $T = |F|^2 / |A|^2$, which we can call the transmission of this barrier, in terms of k , κ and L_b . (Hint: work from the right of this problem back towards the left, solving first for the relations of C and D to F at the right interface using the appropriate boundary conditions.)

- (iii) For a barrier height $V_o = 2$ eV, an incident energy $E = 1.5$ eV and a barrier thickness $L_b = 0.25$ nm, calculate the transmission T of this barrier.

3.7.6 Consider a particle of mass m in a semi-infinite one-dimensional well, as shown here:



The potential in area III is infinite. In area I it is zero. In area II, it has a constant value of $V_o = \frac{18\hbar^2}{mL^2}$. The particle has energy $E = \frac{3}{4}V_o$. (Note: we have verified for you that this is an energy eigenvalue for this problem, so you do not have to derive or prove this is an eigenvalue.)

- Write the mathematical form of the wavefunction in each of the three areas, specifying expressions for the magnitude “ k ” of the wavevector in any sine or cosine terms, and for the exponential decay constant “ κ ” in any exponentially decaying terms.
- Use boundary conditions as necessary to write down a form of the wavefunction for all three areas with only one arbitrary “amplitude” constant for the entire wavefunction. Note: in considering boundary conditions, continuity of wavefunction will suffice - you will not need also to consider the continuity of the first derivative (though, incidentally, because this is an energy eigenfunction of the problem, you would find that condition also to be satisfied).
- Normalize this wavefunction.

3.8 Conclusions

In this chapter, we have progressively introduced many of the core features of the quantum mechanical view of the world. We started from the early history and continued through the introduction of apparently strange and counter-intuitive ideas, such as wave-particle duality, and the proposal that the electron was also a wave. We were forced to these ideas because otherwise we could not construct models of light and matter that agreed with observations. With the proposal of Schrödinger’s wave equation and the equivalent Heisenberg matrix form, these and other pioneers of quantum mechanics were able to construct a model of the hydrogen atom that agrees remarkably

with experiments and that forms the basis of modern chemistry. We will look more at the quantum mechanics of atoms below.

We have seen that a simple example problem, the particle in a box, introduces a wide range of the phenomena of quantum mechanics, including discrete energy levels. The analysis of the behavior of waves has helped us introduce an uncertainty principle. The double-slit experiment exposed many more unusual ideas in quantum mechanics, including the ontological nature of quantum mechanical “particles” and “waves”, and the difficult issue of the measurement problem in quantum mechanics. Finally, we were able to introduce a simple model of the phenomenon of tunneling, which is an effect seen routinely in modern electronic devices.

With this background we are now ready to examine some of the consequences of quantum mechanics in atoms and materials, and to add a few more quantum mechanical concepts.

Chapter 4

Particles, atoms, and crystals

4.1 The hydrogen atom

The hydrogen atom is the simplest of all atoms. It consists of only one electron and only one proton. Despite its simplicity, many of the features we find there are retained, at least approximately, as we work with more complex, multi-electron atoms. So, understanding the hydrogen atom gives us a very good start on understanding atoms, chemistry, and matter generally.

Fortunately, the hydrogen atom is also a quantum mechanical problem that can be solved essentially exactly and completely using Schrödinger's wave equation. There is no particular mathematical difficulty in that solution; it is a straightforward application of standard mathematics of partial differential equations in three dimensions. It does, however, take a fair amount of algebra to go through that solution completely from first principles, so we will not do that here¹.

Even if we do not formally complete the derivation, we can usefully introduce some of the core ideas that allow us to simplify the problem and separate it into different parts. We can also understand both the nature of these solutions and the overall behavior of the results. That understanding will give us a useful conceptual basis for seeing just how all the physical material round about us is put together.

4.2 Center-of-mass behavior

Because the hydrogen atom consists of both an electron and a proton, the first difference in the quantum mechanics compared to a simple particle in a box is that we have to deal

¹ See, for example, Chapter 9 and Chapter 10 of *Quantum Mechanics for Scientists and Engineers* by D. A. B. Miller (Cambridge 2008)

with two particles. Fortunately, there is a way of recasting this problem to give two separate quantum mechanical problems, each of which is like the quantum mechanics of just a single particle again. One of those problems involves considering the combination of the electron and the proton as one particle – essentially the atom considered as a whole – that has a mass M given by the sum of the electron and proton masses. This “entire atom” just behaves like a free particle with that total mass M in its own Schrödinger equation. That equation describes how the whole atom behaves quantum mechanically as a single particle, just as the Schrödinger equation for an electron describes how the electron behaves. We will not discuss that equation further here, in part because we have already dealt with such behavior.

The other quantum mechanical problem is the one that gives the “internal states” of the atom. Classically, for a single planet orbiting a sun, it is not quite correct to say that the center of the orbit is the center of the sun. Actually, the orbit is around the “center of mass” of the combination of the planet and the sun². If the sun is very much more massive than the planet, then we might not notice this effect because the center of mass will be close to the center of the sun, so the radius of the orbit of the sun will be small, likely smaller than the sun itself. The radius of the planet’s orbit will likely be very much larger because it is very much lighter. But, to be correct, both the planet and the sun are orbiting around this “center of mass”.

We can work through the mathematics of this “center of mass” behavior, either classically and quantum mechanically; the final equations we write down, technically for the relative motion of the planet and the sun or the electron and the proton, look like those for a planet (or an electron) orbiting round the center of the sun (or the proton), but in which we use a “reduced mass” for the planet (or the electron).

In the standard formula we get from this “center-of-mass” analysis, the reduced mass μ , is given by

$$\frac{1}{\mu} = \frac{1}{m_e} + \frac{1}{m_p} \quad (4.1)$$

or equivalently

$$\mu = \frac{m_e m_p}{m_e + m_p} \quad (4.2)$$

Using the electron mass $m_e \approx 9.109\,382\,91 \times 10^{-31}$ kg and the proton mass

$$m_p \approx 1.672\,621\,777 \times 10^{-27} \text{ kg} \quad (4.3)$$

the reduced mass has the value

² The position $\mathbf{R}$ of the center of mass for two objects of mass m_e and m_p that are at positions $\mathbf{r}_e$ and $\mathbf{r}_p$ respectively is $\mathbf{R} = (m_e \mathbf{r}_e + m_p \mathbf{r}_p) / (m_e + m_p)$, which is a kind of “weighted average” position. If we were thinking about masses on the two ends of a light bar or a seesaw, the center of mass position would be the position along the bar at which it would be balanced, or equivalently the point where we could put the pivot or fulcrum of the seesaw to get it to balance.

$$\mu \approx 9.104\,424\,485 \times 10^{-31} \text{ kg} \quad (4.4)$$

Because the proton mass is very much larger (by a factor of ~ 1836) than the electron mass, this reduced mass is close to the electron mass. Also, as a practical matter, we usually refer to our resulting solutions of the hydrogen atom problem as being the electron orbitals, though strictly, as we said, they are functions of the relative position of the electron and the proton, not the absolute electron position. From now on, we also will just call our solutions the electron orbitals, with this minor formal understanding.

Quiz questions

Q4.2.1 The electron's mass is

- a) $1/1836$ of a proton's mass
- b) 1836 proton masses
- c) About the same as a proton's

Q4.2.2 Suppose, for the purposes of this question, we think about the electron and proton as if they were classical “point” particles, with definite positions, and attracted to one another by the usual Coulomb attraction of oppositely charged particles. Choose the correct answer.

- a) The electron orbits round a stationary proton
- b) The proton orbits round a stationary electron
- c) The electron and proton each orbit around a center of mass that is very close to the electron position
- d) The electron and proton each orbit around a center of mass that is very close to the proton position

Q4.2.3 Suppose that, instead of considering a hydrogen atom, which has an electron and a “nucleus” that is just a proton, we consider deuterium, which still has just one electron but it has a nucleus that is a proton and a neutron, tightly bound to one another. Such a nucleus still has the same electrical charge as the proton, but has a mass that is approximately twice as large. Consider now the “reduced mass” that is used in the separation of this problem into center of mass coordinates. Choose the correct answer. This reduced mass is...

- a) Larger than the reduced mass used for the hydrogen atom
- b) Smaller than the reduced mass used for the hydrogen atom
- c) The same as the reduced mass used for the hydrogen atom

4.3 The hydrogen atom Schrödinger equation

Potential energy

The electron and the proton have equal and opposite charges, so both charges are of same size – that is, of magnitude $e \approx 1.602\,176\,565 \times 10^{-19} \text{ C}$ – but the electron has

charge $-e$ and the proton has charge $+e$. If we have charges like this separated by a distance r , then they will have an electrostatic “Coulomb” attraction force between them. Like the gravitational force, this is an “inverse square” force – that is, $\propto 1/r^2$ where r is the distance between the two charges.

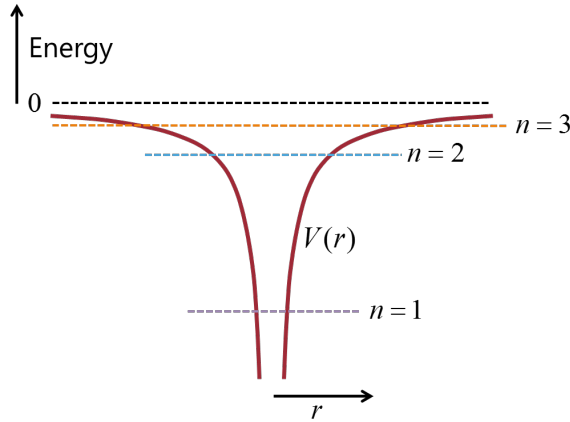


Fig. 4.1. Sketch of the Coulomb potential, $V(r)$, and the corresponding positions of the energies corresponding to values of the principal quantum number $n = 1$, 2, and 3 from the solutions of the hydrogen atom. The chosen zero for energy is shown as the top, dashed line. All the shown energies are negative, corresponding to the attractive nature of the Coulomb potential for the electron and the proton and to the “bound” nature of the states for the various values of n .

If we integrate force times distance, conventionally starting with the two attracting bodies arbitrarily far apart, then we can calculate the potential energy, which then becomes $\propto 1/r$. For the electron and the proton separated by a distance r , this becomes

$$V(r) = -\frac{e^2}{4\pi\epsilon_0 r} \quad (4.5)$$

where

$$\epsilon_0 = \frac{1}{4\pi \times 10^{-7} c^2} \approx 8.854187817... \times 10^{-12} \text{ F/m} \quad (4.6)$$

This potential energy, sketched as the solid curve in Fig. 4.1, is a negative number because, since the electron and the proton are attracting one another, we would need to put energy in to pull them apart. This energy value is relative to the potential energy they would have if they were arbitrarily far apart (so $V(r \rightarrow \infty) \rightarrow 0$); this chosen “zero” of energy is shown as the dashed line nearest to the top of Fig. 4.1.

Relative motion wavefunction 10b

Using the reduced mass and this potential energy, we are ready to write the Schrödinger equation for the “electron” orbitals, that is, for the relative motion wavefunction, which we will call $U(\mathbf{r})$,

$$\left(\frac{-\hbar^2}{2\mu} \nabla^2 - \frac{e^2}{4\pi\epsilon_0 r} \right) U(\mathbf{r}) = EU(\mathbf{r}) \quad (4.7)$$

Note this wavefunction is a function of the vector relative position, $\mathbf{r}$, of the electron and the proton (that is, $\mathbf{r}$ is the electron position vector relative to the proton).

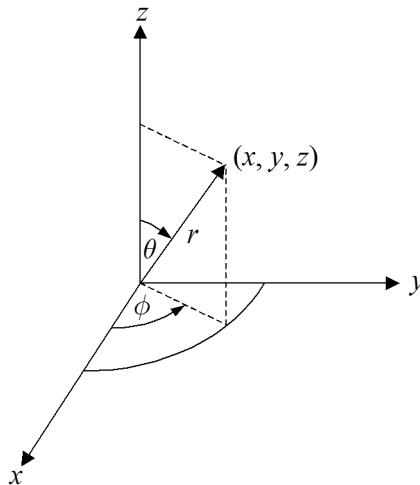


Fig. 4.2. Spherical polar coordinates, r , θ , and ϕ related to the conventional Cartesian x , y , and z coordinate directions.

When we examine this equation in more detail, we find that the solutions of it can be written as the product of two functions, one of which, $R(r)$, only depends on the separation of the electron and the proton, and the other of which, $Y(\theta, \phi)$, only depends on the angles, relative to some chosen directions. That is, we can write

$$U(\mathbf{r}) = R(r)Y(\theta, \phi) \quad (4.8)$$

The coordinate system we use here is like one we could use to describe the position of a satellite, relative to the center of the earth. There would be a distance, r , from the earth's center, and two angles³, one related to the “latitude”, which is the “polar” angle θ , and the other related to the “longitude”, which is the “azimuthal” angle ϕ . This is the “spherical polar” coordinate system, and the formal definitions⁴ can be seen from Fig. 4.2.

³ ϕ corresponds with longitude if we take the angles going East from the meridian (which is a circle centered on the origin and touching the z and x axes) round in a complete circle (that is, to 360°). Latitude is conventionally expressed relative to the equator, but the angle θ is expressed relative to the “North” pole (which is on the positive the z axis), so the “North” pole is at 0° , the equator is at 90° , and the “South” pole is at 180° .

⁴ The choice of θ to represent the “polar” angle (i.e., relative to the North pole or the z axis) and ϕ to represent the “azimuthal” angle (i.e., in the equatorial plane, relative to the x axis), as shown here, is the convention commonly used in physics. In mathematics, it is also common to use the opposite convention, with ϕ being the polar angle and θ being the azimuthal angle.

An expression such as Eq. (4.8), where we can rewrite a function of many variables as a product of functions each of fewer variables, is called a “separation” of the solution or a “separation of variables”⁵. The fact that we can “separate” the solution in this way is because the potential in Eq. (4.5) that we use in our Schrödinger equation (4.7) does not depend on angle at all. Such a potential is called a “central potential” – it only depends on the distance from the center. This is important because, even for more complicated atoms, which may have many protons in the nucleus of the atom and correspondingly many electrons, the potential seen by a given electron can still be approximately independent of angle. Hence those solutions for more complicated atoms still have similar angular shapes as the solutions to the hydrogen atom, at least approximately. This is behind the whole description even of complicated atoms in terms of “orbitals” that use the same nomenclature as we have for the hydrogen atom.

Quiz questions

Q4.3.1 The potential energy $V(r)$ used in the hydrogen atom problem is conventionally written so that it tends to zero as r tends to infinity. With this convention, choose the correct statement below. (Note: Here we use the meaning that “increasing” means becoming a larger positive number or a smaller negative number. This is like saying that, on a vertical graph axis, “increasing” a value means moving it up on the graph, regardless of whether it is below or above the horizontal axis.)

- a) The potential energy increases as the nominal distance r between the electron and the proton reduces, and is proportional to $1/r$.
- b) The potential energy decreases as the nominal distance r between the electron and the proton reduces, and is proportional to $1/r$.
- c) The potential energy increases as the nominal distance r between the electron and the proton reduces, and is proportional to $1/r^2$.
- d) The potential energy decreases as the nominal distance r between the electron and the proton reduces, and is proportional to $1/r^2$.

Q4.3.2 Consider the Earth as if it were a perfect sphere and presume that we are describing our position on the Earth using spherical polar coordinates with the direction from the center of the Earth to the North Pole corresponding to the positive z direction. State for each of the following whether it is true or false.

- a) As I walk north from the equator towards the North pole, I am decreasing the magnitude of my polar angle.
- b) As I walk north from the equator towards the North pole, I am increasing the magnitude of my azimuthal angle.

Q4.3.3 Suppose now we presume that the Greenwich observatory in London lies in the x - z plane as we have drawn the spherical coordinates. State for each of the

⁵ Sometimes this kind of product form is called an “ansatz”, especially when we initially propose it as a kind of educated guess before verifying it does actually work.

following whether it is true or false. [Note carefully that we said the x - z plane here, not the x - y plane.]

- a) As I walk east from Greenwich, my polar angle is decreasing.
- b) As I walk east from Greenwich, my azimuthal angle is increasing.

Problem

4.3.1 In the Schrödinger equation for the hydrogen atom, the potential energy is

$$V(r) = \frac{e^2}{4\pi\epsilon_0 r}$$

where the various fundamental constants have their usual meanings, and r is the quantity representing the distance between the electron and the proton. True or false?

4.4 Spherical harmonics

The solutions for the angular part in Eq. (4.8), that is for the function $Y(\theta, \phi)$, are called “spherical harmonics”. These functions can look somewhat intimidating at first glance. Fortunately, we can understand what these functions are like without performing the full solution here.

Mathematics of spherical harmonics

The formula for these functions can be written out completely. Explicitly, in quantum mechanics we use the form

$$Y_{lm}(\theta, \phi) \propto P_l^m(\cos \theta) \exp(im\phi) \quad (4.9)$$

where we are indexing the different possible spherical harmonic functions with the integers l and m . Here, the functions $P_l^m(x)$ are the associated Legendre functions, whose detailed definitions and behavior can be found in standard mathematical reference works and in some mathematics programs. These functions are real, so the only complex aspects of the spherical harmonics come from the $\exp(im\phi)$ part of the functions. The precise details of the associated Legendre functions are not particularly important here, so we omit them, though we can understand their important qualitative behavior.

The most important aspect of these spherical harmonics concerns the integers l and m associated with them. When we are considering the angular behavior of the possible solutions of the hydrogen atom, there is one important underlying principle: essentially, if we go round in a circle, we have to get back to where we started. Equivalently, any function $Y(\theta, \phi)$ that is to be a solution for the spatial wavefunction of the hydrogen atom must be periodic in each of the angles θ and ϕ ; that is, we require

$$Y(\theta, \phi) = Y(\theta + 2\pi, \phi) = Y(\theta, \phi + 2\pi) \quad (4.10)$$

Otherwise, the function $Y(\theta, \phi)$ would not be “single-valued” as a function of the physical angles. For any function whose argument is $\cos \theta$, we are automatically guaranteed that the function is periodic in this way for the angle θ because $\cos \theta$ is periodic in this way, so $P_l^m(\cos \theta)$ will be periodic in this way, no matter what the functional form of the associated Legendre functions happens to be.

For the function $\exp(im\phi)$ to be periodic in this way with angle ϕ , m must be an integer. Explicitly, we would require

$$\exp(im\phi) = \exp[im(\phi + 2\pi)] = \exp(2\pi im)\exp(im\phi) \quad (4.11)$$

which in turn requires

$$\exp(2\pi im) = 1 \quad (4.12)$$

which requires that m is an integer.

The detailed solution of the equations leading to the result Eq. (4.9) then also requires that l is zero or a positive integer, that is

$$l = 0, 1, 2, 3, \dots \quad (4.13)$$

and also that m lies between $-l$ and $+l$, that is

$$-l \leq m \leq l, \text{ with } m \text{ integer} \quad (4.14)$$

In our full hydrogen atom solutions, l and m are going to become “quantum numbers” that describe the solutions. In addition, the full solution to the hydrogen atom introduces one other quantum number, n , that is associated with the allowed energies of the hydrogen atom, in the same way as we saw before in the Bohr model of the atom. So we have three quantum numbers:

n is the principal quantum number

l is the orbital quantum number, or the angular momentum quantum number

m is the magnetic quantum number, or the azimuthal quantum number

Visualizing spherical harmonics

Spherical harmonics can arise in classical problems as well. One good example is the vibration modes of a spherical shell – a thin shell in the shape of a sphere, like an inflated balloon or ball. In the case of classical problems, we only have real amplitudes, of course. In that case, we could have two sets of spherical harmonic functions, one in which we replace the $\exp(im\phi)$ in Eq. (4.9) with $\cos(m\phi)$, which are functions that are just the real part of our complex spherical harmonic functions of Eq. (4.9), and another in which we replace the $\exp(im\phi)$ with $\sin(m\phi)$, which are functions that are just the imaginary part of our complex spherical harmonic functions of Eq. (4.9). These two different cosine and sine sets of spherical harmonic functions are really just the same functions but rotated by $90^\circ/m$ in the azimuthal (equatorial) plane, so we only need to understand the form of one of these sets to understand the entire set of spherical harmonic functions.

In the classical case, instead of letting m run from $-l$ to $+l$, we simply choose it to be zero or positive, and use both the cosine and sine forms. This gives us the same total number of functions⁶. Incidentally, when spherical harmonic functions are plotted, even in discussions of quantum mechanical problems such as the hydrogen atom, it is typically these real forms that are shown. For $m > 0$, the sine and cosine forms have the same shapes (with the sine versions just being rotated around the “vertical” axis compared to the cosine versions), so we only need to plot one form to see the shapes.

The different modes of vibration of a spherical shell can be understood from Fig. 4.3. As with other oscillating or vibration modes, everything that is vibrating in such a mode is vibrating at the same frequency. The spherical harmonics here are going to tell us the amplitudes of those vibrations, in the same way that a function like $\sin(n\pi z / L)$ would tell us the amplitude of the oscillation of a wave on a string tied between two posts a distance L apart in the z direction.

In the case of the wave on a string, for (integer) n larger than 1, there are zeros in the oscillation at specific points. For example, for $n = 2$, there is a zero in the middle of the string. Such zeros for oscillations can also be called nodes. In the case of the vibrations of a spherical shell, instead of just specific points where there is no vibration, we have complete circles round the sphere where there is no vibration, even though the rest of the spherical shell is vibrating. We can call these specific circles “nodal circles”.

Just as with the wave on a string, where there is no node along the length of the string (other than the ends) for the “lowest” mode, then for the spherical shell, there is one mode, called the “breathing mode”, that has no nodal circles. This mode is as if we were inflating and deflating the balloon slightly. So the angular amplitude of the oscillation is the same in all directions. This “breathing” mode is indicated by the sphere shown for $l = 0$, $m = 0$ in the top left of Fig. 4.3. The next set of modes have one nodal circle. For example, for the one in the top right of Fig. 4.3, indexed with $l = 1$ and $m = 0$, the “equator” of the sphere is a ring on which there is no vibration at all. Just as would be the case on either side of a node for the case of waves on a string, the oscillation is in the opposite sense on one side of the nodal circle compared to the other. So for this sphere with one nodal circle round the equator, it is as if the top hemisphere is inflating while the bottom hemisphere is deflating during one half of the oscillation cycle and the opposite way round during the other half of the oscillation cycle. The situation is similar for the $l = 1$, $m = 1$ case in the middle-left of Fig. 4.3, except it is the “left” and “right” hemispheres that are inflating and deflating oppositely.

As we proceed to the remaining cases in Fig. 4.3, we have more nodal circles (two for each of the remaining cases shown). The $l = 2$, $m = 0$ case is an oscillation that looks like we are alternately “squashing” and “stretching” the sphere along the polar direction; when we squash at the poles, the sphere expands around the equator, and when we stretch along the poles, the sphere shrinks round the equator. The remaining two cases in Fig. 4.3 correspond to adjacent “quarters” of the sphere “inflating” and “deflating”

⁶ When $m = 0$, there is no variation with angle ϕ , and the “sine” spherical harmonic would be zero everywhere, so we do not need to count it, so there is only one function associated with $m = 0$ in this “classical” or “real” case also.

oppositely. Note that, in all these oscillating modes, the corresponding nodal circles, which we could think of as lines drawn on the surface of the sphere, do not move at all.

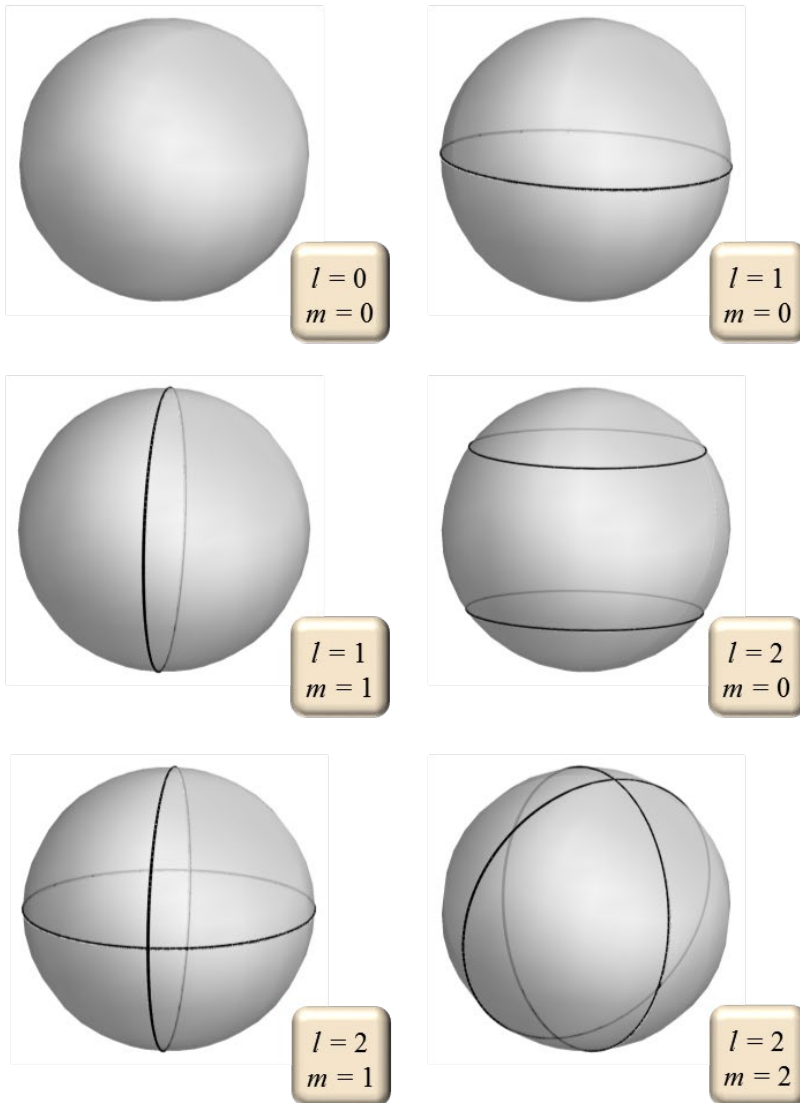


Fig. 4.3. Plot of the nodal circles for the vibration modes of a spherical shell, together with the numbers l and m corresponding to them. Note that l gives the total number of nodal circles, and m gives the number that pass through the “North” and “South” poles of the spherical shell. (The “poles” are tilted and rotated slightly in some of the cases for greater clarity in the drawings.)

As we can see directly from Fig. 4.3, there is a simple behavior emerging for the nodal circles and the quantum numbers l and m

$$l \text{ is the total number of nodal circles} \quad (4.15)$$

$$|m| \text{ is the number of nodal circles through the poles} \quad (4.16)$$

This remains true for all the higher modes of oscillation of the sphere, and is a very useful relation for thinking about spherical harmonics.

Polar plots of spherical harmonics 10d

We can also plot these spherical harmonic functions in another way. They are functions of the two angles, θ and ϕ , so we could plot a surface whose distance from the center was the (real) amplitude of the spherical harmonic. The only subtlety here is that the amplitude of the spherical harmonic function can change from positive to negative as a function of angle, but a distance from the center can only really be positive. So we plot the magnitude of the (real) spherical harmonic function as the distance from center as a function of angle, and mark the different resulting regions as being either positive or negative. The results for four representative cases are shown in Fig. 4.4.

Note that there is a relatively simple conceptual way of getting from the “nodal circles” picture of Fig. 4.3 to the “surface” plots of Fig. 4.4, at least qualitatively. Suppose that each of the nodal circles in Fig. 4.3 is an elastic band on the surface of the sphere, and suppose we now shrink those elastic bands to the point where they completely “pinch” the spherical “balloon”. We need to give the elastic bands one unusual property, which is that they always pinch down to the center of the sphere (we need this, for example, for the $l = 2, m = 0$ case).

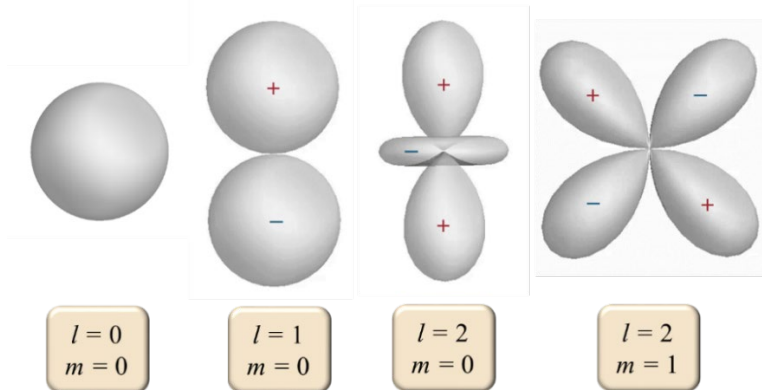


Fig. 4.4. Plots of four of the (real) spherical harmonics, with distance from the center in each case representing the magnitude of the spherical harmonic. The “+” and “-” signs indicate the relative sign of the spherical harmonic for different segments. The plots are not to any specific scales, being meant to represent the form of the functions only.

Hence, we can now qualitatively visualize any of the spherical harmonic functions, without any calculations. First we draw the nodal circles on a sphere, following the rule (4.15) for the total number of nodal circles (i.e., l) and rule (4.16) for the number through the “poles” (i.e., m). We then mark the sphere with opposite signs on opposite sides of each nodal circle. Then we consider those nodal circles to be bands that we then use to pinch the spherical “balloon”, with the one special property that the bands always pinch down to the center of the balloon. With this approach, we can qualitatively construct all of the plots as in Fig. 4.4.

Spherical harmonics and hydrogen atom orbital notation 10e

Historically, the spectral lines observed in atoms were collected into groups, which were given what we could now regard as relatively arbitrary names. Specifically, these groups were called

- “sharp” (s)
- “principal” (p)
- “diffuse” (d), and
- “fundamental” (f)

Now, we associate these terms directly with the l quantum number, as in Table 4.1.

l	0	1	2	3	4	5
notation	s	p	d	f	g	h

Table 4.1. Letter notations for the l quantum number.

Only s, p, d, and f are in common use because the electronic configurations of the “ground” states of all known atoms can apparently be described in terms of only these four. The notation can be extended alphabetically as shown, and this extended notation can be used to label excited states of atoms.

Angular momentum and spherical harmonics

Bohr had originally proposed a quantization of angular momentum in his early model of the hydrogen atom, in units of $\hbar$. Though this initial proposal was not exactly correct in detail, the essence of this idea survives into the model that results from the solution of the Schrödinger equation. In this Schrödinger equation model, the hydrogen atom orbitals have an angular momentum around the polar (z) axis of an amount $m\hbar$.

The reason why the quantum number m is known as the magnetic quantum number is related to the angular momentum. We know from classical electromagnetism that, if we had a current running in a loop of wire and we applied a magnetic field in the direction perpendicular to the loop, then there would be an energy associated with this that was proportional both to the size of the magnetic field we are applying and to the current in the wire. We can understand this because we know that having a current going in a loop itself generates a magnetic field; equivalently, we can think of that loop carrying a current as behaving like a magnet. Not surprisingly, having a magnet in a magnetic field results in some energy. For example, we know the magnet “wants” to orient itself in a specific direction relative to the magnetic field, and we can extract mechanical energy from the system by allowing it to do that – this is how typical electric motors work⁷. If

⁷ We can have motors that work on the basis of electrostatic forces only, and a few actuators do work this way, especially very small micromechanical ones, but most so-called electric motors work on the basis of the forces generated by magnetic fields.

we have in our head the idea of a “Bohr” classical orbital of a point charge going round in a circle, then it is easy to connect the idea of the angular momentum of the electron going round in a circle with the current in this classical loop, and hence we would indeed expect the energy of such a state of a specific angular momentum to be proportional to the angular momentum and to the magnetic field.

Though this classical “Bohr” model of the hydrogen atom is not correct, when we solve the quantum mechanical version of problem of such orbitals in the presence of a magnetic field, we do indeed get a splitting of the energy levels that corresponds to what we would expect from a “Bohr”-like model, based on orbitals having angular momentum $m\hbar$. The splitting of energy levels of atoms with magnetic fields is called the Zeeman effect⁸. Incidentally, one reason we use the complex exponential version of the spherical harmonics, as in Eq. (4.9), with positive and negative values of m , rather than the sine and cosine versions with only non-negative m , is to provide a more convenient description of these magnetic effects. In a magnetic field, the negative and positive m values correspond to energy levels with opposite variations with magnetic field, i.e., there is a change in energy with magnetic field that is proportional to this “negative or positive” m .

A key difference in the results of the Schrödinger equation compared to the “Bohr” model is that the quantum number in the angular momentum is m not n , and it is also quite possible for m to be zero; some of the possible orbitals of the hydrogen atom indeed have zero angular momentum in this sense. To understand the relation between both the orbital angular momentum quantum number l , the magnetic quantum number m , and the principal quantum number n , we need now to look at the solutions for the radial part of the wavefunctions.

Quiz questions

Q4.4.1 For spherical harmonic functions $Y_{lm}(\theta, \phi)$ in the conventional notation we are using, and expressed as a function of the usual θ and ϕ polar and azimuthal angles, state whether each of the following is True or False. No matter what allowed values of l and m we choose for the spherical harmonic function $Y_{lm}(\theta, \phi) \dots$

- a) $Y_{lm}(\theta, \phi) = Y_{lm}(\theta, \phi + 2\pi)$
- b) $Y_{lm}(\theta + 2\pi, \phi) = Y_{lm}(\theta, \phi)$

Q4.4.2 State whether each of the following is a possible allowed combination of l and m for a spherical harmonic function in the form as stated in the text.

- a) $l = 0, m = 0$
- b) $l = 0, m = 1$

⁸ There are also additional splittings of the energy levels with magnetic field that are associated with the spin of the electron, a phenomenon not expected or predicted in the “Bohr” model.

c) $l = 1, m = 0$

d) $l = 1, m = -1$

e) $l = 4, m = 5$

Q4.4.3 In the classical problem of the modes of vibration of spherical shell, and considering the “real” (rather than complex) versions of the spherical harmonics, consider the “nodal circles” – the circles on the sphere that have no amplitude of oscillation – for given such modes. We consider the spherical harmonics, formally in spherical polar coordinates, in which the z axis runs from the South pole to the North pole.

- What is the l -value for a spherical harmonic with a total of 4 nodal circles, three of which pass through the poles?
- What is the m -value for this same spherical harmonic?
- For this same spherical harmonic, how many nodal circles lie along the sphere's equator? (Hint: it has to be either zero or one – there is only one equator on a sphere.)

Q4.4.4 Atomic orbitals are often labelled using a notation using a number for the “principal” quantum number n followed by a letter representing the l value, using the “spdf” notation. State the correct l value for each of the following orbitals labelled this way.

- 1s
- 4f
- 3d
- 2p
- 4s

Problem

4.4.1 For each of the following labels in the “spdf” notation, state the corresponding orbital angular momentum quantum number l .

- d
- g
- j

4.5 Radial solutions and energy levels

When we solve the appropriate separated equation for the radial function $R(r)$ that we introduced in Eq. (4.8), the resulting solution can be written in the form (see Fig. 4.5)

$$R_{nl}(r) \propto \left(\frac{2r}{na_o} \right)^l L_{n-l-1}^{2l+1} \left(\frac{2r}{na_o} \right) \exp \left(-\frac{r}{na_o} \right) \quad (4.17)$$

Here, the expression for the Bohr radius a_o , which is identical to what we would obtain from Bohr's original model, is given by the expression

$$a_o = \frac{4\pi\epsilon_o\hbar^2}{e^2\mu} \cong 0.529 \text{ \AA} = 5.29 \times 10^{-11} \text{ m} \quad (4.18)$$

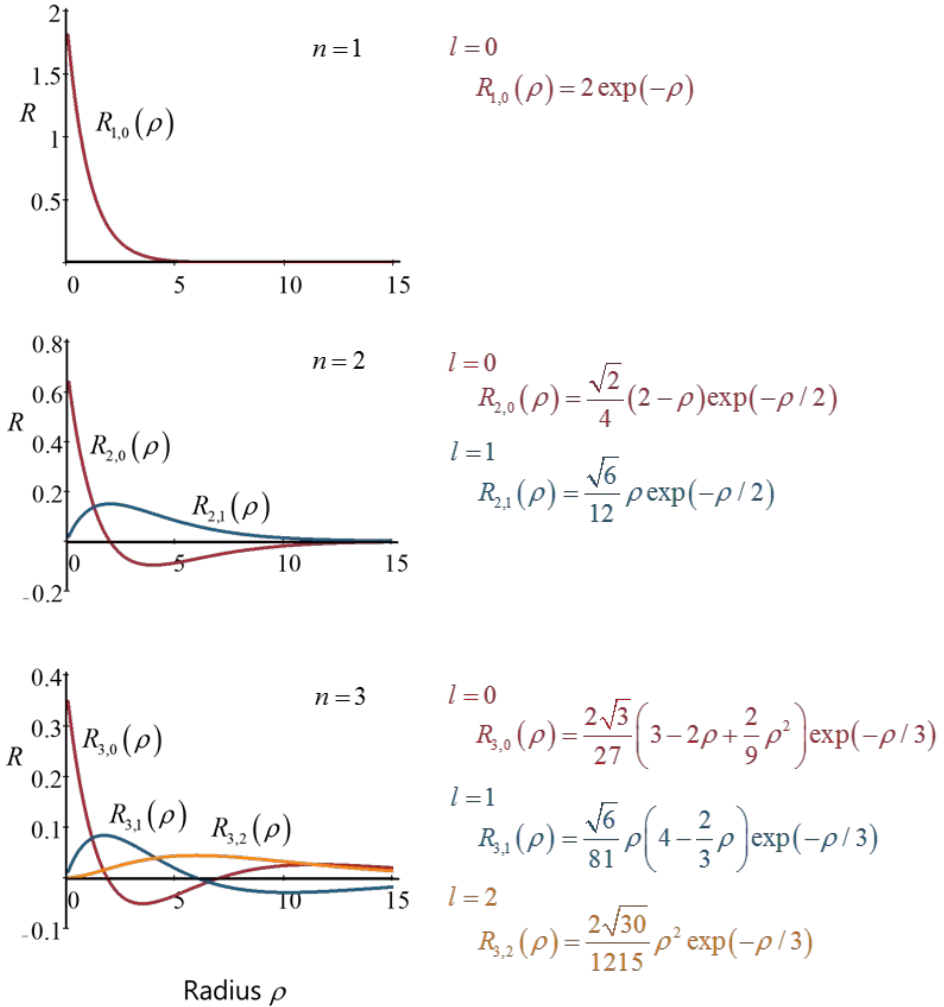


Fig. 4.5. Plots of the radial functions for the hydrogen atom with distance in units of the Bohr radius ($\rho = r/a_o$ where r is the distance from the center and a_o is the Bohr radius).

The functions $L_p^l(s)$ are the associated Laguerre polynomials, whose precise definitions can be found in mathematical references and some mathematical programs.

Two important mathematical requirements for these eigen solutions are that we must have

$$n \text{ is an integer, starting at 1, that is } n = 1, 2, 3, \dots \quad (4.19)$$

$$n \geq l + 1, \text{ or, equivalently, } l \leq n - 1 \quad (4.20)$$

When we substitute this total solution of the form of Eq. (4.8), with our results Eqs. (4.9) and , back into the original Schrödinger eigen equation for the hydrogen atom, Eq. (4.7), the resulting eigen energies – that is, the allowed values for E – are⁹

$$E_{ln} = -\frac{Ry}{n^2} \quad (4.21)$$

where now we can formally write out the formula for the “Rydberg” Ry

$$Ry = \frac{\hbar^2}{2\mu a_o^2} = \frac{\mu}{2} \left(\frac{e^2}{4\pi\epsilon_o \hbar} \right)^2 \approx 13.6 \text{ eV} \quad (4.22)$$

Note one particularly important point: These energies of the hydrogen atom states do not depend on either l or m , but only on n , the principal quantum number. Given the different shapes of the orbitals for different n , l , and m , this is a surprising result. (Actually, hydrogen is unique among the atoms in reliably having this property; for other atoms, the eigenenergies can depend on all the quantum numbers.)

The energies corresponding to $n = 1, 2$, and 3 are sketched in Fig. 4.1.

Again, though the expression Eq. (4.17) may look somewhat daunting, we can understand relatively simply what these functions mean. These radial solutions and the explicit formulas for them are plotted in Fig. 4.5. Note the following attributes of these solutions:

The overall “size” of these wavefunctions becomes larger with larger n .

The wavefunctions generally have an overall exponential decay of the form $\exp(-r / na_o)$, so this exponential decay is slower with r for larger n .

These radial functions have $n - l - 1$ zeros (not including any zero at the center ($r = 0$), or any “asymptotic” zero as $r \rightarrow \infty$)

These zeros are from the roots of the corresponding polynomial functions.

This completes our mathematical solution of the hydrogen atom problem.

Quiz questions

Q4.5.1 Taking the energy origin as usual for the hydrogen atom (which corresponds to the potential energy tending towards zero at very large distance), choose the correct energy for each of the following hydrogen atom orbitals:

- a) 3s: $-13.6(\text{eV})/9$ or $-13.6(\text{eV})$
- b) 3p: $-13.6(\text{eV})/9$ or $-13.6(\text{eV})/4$

⁹ Note this expression for the hydrogen atom energies is identical to the form originally deduced from Bohr from his earlier semiclassical model.

Q4.5.2 For each of the following hydrogen atom orbitals, state how many zeros there are as a function of radius from the center (this will, of course, be an integer number). (Do not include any zero at the center, and do not include any (asymptotic) zero at infinity).

- a) 3s
- b) 2s
- c) 3d

Q4.5.3 For each of the following, state whether it is an allowed combination of n , l , and m for describing possible orbitals (i.e., eigen solutions) of the hydrogen atom.

- a) $n = 0, l = 0, m = 0$
- b) $n = 2, l = -1, m = 0$
- c) $n = 2, l = 3, m = 0$
- d) $n = 3, l = 1, m = 2$
- e) $n = 4, l = 2, m = -2$

Problems

4.5.1* The hydrogen atom is formed from an electron and a proton. It is also possible to form a “positronium atom”, which is made from an electron and a positron. The positron is the “antiparticle” to the electron; it has the same mass, but its charge is equal but opposite to that of the electron (i.e., it is $+e$ where the electron charge is $-e$). The behavior of the positronium atom can be calculated in exactly the same way as that of the hydrogen atom, but the resulting size and energies are different because of the different mass of the positron compared to the proton.

- (i) Write down an expression in terms of fundamental constants and a principal quantum number n for the energies of the “internal states” of the positronium atom.
- (ii) Evaluate the magnitude of the “binding energy” of the positronium atom (which is the same as the magnitude of the energy of the $n = 1$ level), and give your answer in electron-volts.
- (iii) Write down an expression for the “Bohr radius” of this positronium atom.
- (iv) Evaluate this “Bohr radius” for the positronium atom, and give your answer in nanometers.

4.5.2 State for each of the following proposed combinations of quantum numbers (n, l, m) whether it corresponds to a “legal” possible set of such quantum numbers to represent a possible energy eigenstate of the hydrogen atom (technically the “internal states” of the hydrogen atom)? (Here we use the standard notation for the principal quantum number (n), the orbital angular momentum quantum number (l), and the magnetic quantum number (m)).

- (i) (1,0,0)
- (ii) (2,1,2)
- (iii) (1,2,2)
- (iv) (2,1,0)
- (v) (3,2,-3)
- (vi) (4,3,2)
- (vii) (2,0,-1)
- (viii) (-2,-1,0)

4.5.3 (Note: for this question, consider that electrons can have two different spins, corresponding respectively to “magnetic spin quantum numbers” of $m_s = 1/2$ and $m_s = -1/2$. So, such electrons of different spins have to be counted when counting total numbers of electrons. Presume, though, that different “orbitals” correspond just to different n , l , and m quantum numbers.)

- (a) How many possible orbitals are there for $n = 5$?
- (b) How many electrons can occupy all of the $n = 5$ orbitals?
- (c) How many electrons in an atom can have the quantum numbers $n = 4$, $l = 2$?
- (d) Which of the following combinations of quantum numbers is (are) not allowed for an electron in a hydrogen atom?

	n	l	m	m_s
(i)	2	0	0	-1/2
(ii)	3	-2	-2	1/2
(iii)	4	3	1	-1/2
(iv)	4	4	-1	1/2

- (e) For $n = 3$, $l = 0$, at what radius or radii would the radial wavefunction become zero? (Give any numerical answers in angstroms.)

4.5.4* What is the energy separation, in electron-volts, between the hydrogen atom states corresponding to the following pairs of orbitals? (Note: here we are using a notation “ $n a$ ” where the number n is the principal quantum number, and the letter a corresponds to the “s, p, d, f, ...” notation, so $5p$ would correspond to $n = 5$ and $l = 1$.)

- (i) $3d$ and $2s$
- (ii) $6g$ and $4f$
- (iii) $2p$ and $2s$

4.5.5 Calculate the energy in eV and the wavelength of the photon emitted if an electron in hydrogen is transitioned from the initial orbital to the final orbital listed below by emitting the photon. (Presume the photon energy has to equal the energy separation between the energies of the orbitals.)

- (a) Initial: $n = 6$, $l = 2$, $m = 1$. Final: $n = 2$, $l = 1$, $m = 0$.
- (b) Initial: $n = 3$, $l = 2$, $m = -2$. Final: $n = 1$, $l = 1$, $m = -1$.

(c) Initial: $n = 5, l = 3, m = 2$. Final: $n = 3, l = 2, m = 1$.

4.6 Overall form of the hydrogen atom orbitals 11b

To see the complete form of the hydrogen atom orbitals, we need to put together the spherical harmonics as in Eq. (4.9) and the radial functions as in Eq. (4.17). To get a sense of what the full behavior looks like, we can plot cross-sections through the middle of the atom of the probability density for finding an electron at a given distance from the proton.

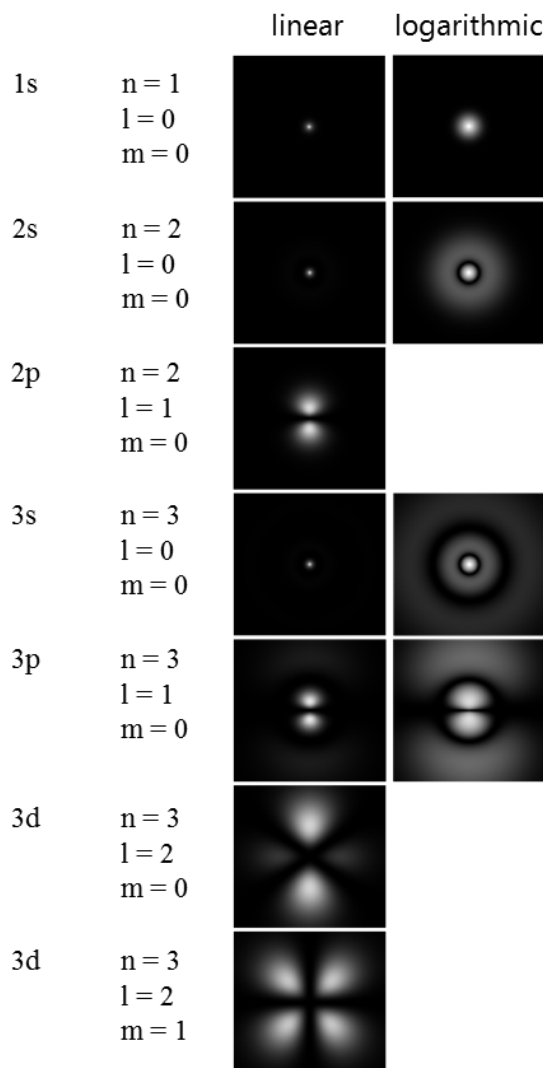


Fig. 4.6. Cross-sections in the $x - z$ plane of the probability density for various hydrogen atom orbitals, shown as “brightness”. To make the behavior clearer for various orbitals, some of these are plotted on a logarithmic “brightness” scale.

These orbitals are shown in Fig. 4.6. As in Fig. 4.4, these are plotted with the z axis in the vertical direction¹⁰. For the s orbitals, these are so small on this scale and their probability density falls off so rapidly with distance that the various radial zeros are not obvious, so we plot these also on a logarithmic “brightness” scale. Similarly for the $3p$ orbital, the logarithmic scale is essential to see the radial zero.

Problem

4.6.1 Suppose we consider hydrogen atom orbitals oriented with respect to a particular polar axis, which we can also call the z axis. We can call the resulting x - y plane the “equatorial” plane. We will call a hydrogen atom orbital “symmetric” with respect to the equatorial plane if the wavefunction has the same value at some point at a height z_0 above the equatorial plane as it does at the corresponding point below the equatorial plane (that is, with the same “ x ” and “ y ” coordinates, but with a z coordinate of $-z_0$). If the orbital has the opposite sign for the corresponding point below the equatorial plane, we will call it “antisymmetric” with respect to the equatorial plane. For each of the following possible orbitals of a hydrogen atom, state whether it is symmetric or antisymmetric with respect to the equatorial plane. (We are using the same notation as for question 4.5.2 above.)

- (i) (4,2,0)
- (ii) (2,1,-1)
- (iii) (2,1,1)
- (iv) (3,1,0)
- (v) (3,2,-2)

¹⁰ So far, we have discussed spherical harmonics only with the z direction being vertical. But, the reader might ask, what sets that vertical direction? Is one direction in space different from the others, and if so why? The answer here is that there is no such special direction (at least unless we do something like applying an electric or magnetic field along some direction). It does not actually matter what direction we choose for z here, though it appears here that we have to choose one so we can solve the problem. The solution to this paradox lies in the fact that this is what is known as a “degenerate” eigen problem. A degenerate eigenvalue problem is one for which there are multiple different possible solutions for the same eigenvalue. Such problems typically arise in situations that have some symmetry. Here, for a given value of l , there is an infinite number of different ways we could write down solutions, each corresponding to a different direction for z . The key point, though, is that if we choose a particular “first” z direction, the solutions for any other choice of z direction can be written down as orthogonal linear combinations of our “first” set of solutions (and *vice versa*). To solve this problem and write down this set of solution functions, we need to choose a direction for z , but we can easily (at least in a fundamental mathematical sense) write down the solutions we would have found with any other choice of direction for z in terms of the specific solutions we devised for our specific “first” choice of the direction for z . Though we need to choose a direction to solve the problem for the first time, we have essentially solved it for any direction of z , and any such choice of z is an equally “good” choice as far as the problem is concerned.

4.7 Spin, electrons, and Pauli exclusion

By the early 1920's it was becoming clear that it was possible to provide some rules for the grouping of different kinds of atoms, based on electrons in “shells”, for example of 2 electrons, 8 electrons and 18 electrons. Wolfgang Pauli realized in 1924 that if he took 4 quantum numbers, three of which correspond to the n , l , and m that we have now introduced for the hydrogen atom, and one more (which we now refer to as electron spin) that could only have two values, then he could explain the categorization of the electron behavior of the various atoms; he proposed that there could be no more than one electron per “state”. By a “state” here we mean something with a unique set of quantum numbers. The restrictions he had on the n , l , and m are essentially the same as we have introduced above for their allowed values as in Eqs. (4.19), (4.20), (4.13), and (4.14). This notion that there can only be one electron in each such “state” is called the Pauli exclusion principle. The two different “values” of the spin for an electron are often referred to as “spin-up” and “spin-down”.

The idea of calling this property “spin” comes from the fact that, at least qualitatively, this property of spin correlates with the idea of a charged mass spinning on its own axis. Such spinning would mean the mass has angular momentum. A charged classical body of some finite size spinning on its axis would be the equivalent of a current loop, and hence would have magnetic moment. The spin of the electron does indeed correspond to an angular momentum, and the electron does also have a “magnetic moment” – that is, it does behave like a very small magnet, as we can see in the splitting of electron energy levels with applied magnetic fields.

The spin of the electron does not, however, correspond quantitatively with any such classical spinning body, and it has some surprising properties quantum-mechanically. One key property is that electron spin is associated with an angular momentum of magnitude $\hbar/2$. So, instead of an orbital angular momentum quantum number l that takes integer values, we have a spin quantum number s for the electron where

$$s = \frac{1}{2} \quad (4.23)$$

Though this approach does work, it becomes clear at this point that the electron spin is not associated with a spatial wavefunction corresponding to an orbit, because this non-integer value means any such spatial wavefunction would not get back to where it started on going round in a circle. Spin is therefore a more abstract property that we cannot completely visualize in a spatial way.

We can, however, define a magnetic spin quantum number, m_s , analogously to the (orbital) magnetic quantum number m and its relation to l , that is,

$$-s \leq m_s \leq s \quad (4.24)$$

which means that we can have

$$m_s = \frac{1}{2}, \text{ known as “spin-up”} \quad (4.25)$$

and

$$m_s = -\frac{1}{2}, \text{ known as "spin-down"} \quad (4.26)$$

With this definition, we can still write the angular momentum associated with these spin states as being $m_s \hbar$, however, and this appears to work in explaining many behaviors.

At the time when Pauli proposed it, spin and the Pauli exclusion principle were introduced for purely empirical reasons in trying to understand the occupation of electron orbitals and without any deeper justification. In 1928, Paul Dirac constructed an equation for the electron that was correct within special relativity. (The Schrödinger equation is not relativistically correct, being based on non-relativistic classical mechanics.) When he did that, the property of spin emerged as a necessary requirement of his equation¹¹. The “explanation” why such “spin $\frac{1}{2}$ ” particles¹² like the electron must obey the Pauli exclusion principle comes from the “spin-statistics” theorem, which is based on relativistic quantum mechanics.

Because the Pauli exclusion principle prevents multiple electrons from occupying the same states, it has a major role in determining the volume occupied by matter and in giving solid materials the mechanical properties they have.

Problem

- 4.7.1 For an electron, there is only one possible value for the magnetic spin quantum number, m_s , which is $\frac{1}{2}$. True or False?

4.8 Many-electron atoms

Potentials and wavefunctions

Hydrogen is the simplest of all the atoms. With only one proton, and hence only one positive charge on its nucleus, it is said to have atomic number 1. Atomic number, Z , is the number of protons in the nucleus, which is the same as the number of electrons in an atom that is not ionized in any way. Armed with the Pauli exclusion principle, we can start to try to understand many electron atoms by looking at the progressive filling of orbitals by electrons.

¹¹ Dirac’s equation also simultaneously predicted the existence of the positron, the “anti-particle” of the electron, a point later clarified by Dirac in 1931. At that time, there was no reason to consider any such “anti-particles”, which can annihilate one another, turning into energy of some kind (such as photons). Now anti-particles are a core part of elementary particle physics. The positron was conclusively discovered in 1932 by Carl Anderson.

¹² The proton and the neutron also are spin-1/2 particles, like the electron, and also obey the Pauli exclusion principle.

Such many-electron atoms can be complicated because the electrons all interact with one another through their Coulomb repulsion. A useful approach is to presume that each electron sees an average background potential energy from all the other electrons, together with the potential from the nucleus¹³. This is an approximation – it is neglecting much possible detail of interactions between individual particles. If we set such details aside, however, we can understand roughly how such many-electron atoms behave.

Such approaches using average background potentials can generally work well enough, and simplify the problem so much, that they are widely used, not just for many-electron atoms but also for many other complicated systems, like molecules or solid materials. In these approaches, the effective average background potential typically has to be determined iteratively; for example, we could guess a net potential energy from all the electrons and the nucleus, calculate the resulting charge density from all the electron states presuming that potential, and then recalculate the net potential energy using the new estimate for the charge density, and so on. Often such methods compare the results of the calculations with observations, such as measured energy levels in the actual atom or material, and adjust parameters in the proposed potential to try to get the model calculations to agree with the experimental results. It might seem on the face of it that there is no point to such a method – why would we set up a model just to give us answers such as energy levels we already know? Such a model, however, can then be used to estimate other aspects, such as wavefunctions, that can usefully predict other phenomena, like chemical bonding or optical properties.

One other major simplifying approximation we can make for atoms is to presume that this average background potential is approximately a “central” one – that is, approximately, it does not depend on angle; that allows us to continue to use spherical harmonics to describe the angular behavior. The problem still “separates” into one with the spherical harmonic angular behavior, and some radial function. The radial function will now be different from that of the hydrogen atom because the potential seen by one electron is now the sum of the potential from the charge of the nucleus and from the charge of the other electrons in the atom. We expect, however, that there will be a set of radial functions that we can still index by some quantum number n , and that those functions will have some behaviors that are broadly similar to those of the hydrogen atom radial functions, for example, in the number of “zeros” and in the relation between the allowed values of l and m and the quantum number n , as in Eq. (4.20).

Filling shells in atoms 12b

For an atom with a large atomic number Z , we can guess that there will be two electrons (one for each spin) in the two 1s-like states (of opposite spin) that are very tightly bound near the nucleus because they will see the very strong attraction from the nucleus with its large number Z of protons. This pair of electrons in that state will form a kind of full “shell”. An electron in a state that is mostly outside that shell (which will have to be one of higher n because the inner shell is “filled”) will see a lower effective attractive charge because the positive charge of the nucleus will be partly “screened” by those

¹³ An approach like this is sometimes called a Hartree-Fock method.

inner two electrons partially cancelling the charge of the nucleus. Then we might expect to consider 2s-like and 2p-like states for the next electrons we add into this $n = 2$ “shell”; because of this screening by the 1s shell, electrons in the states in the second “shell” will see a somewhat lower effective positive charge attracting them towards the center. The 2p-like states are in general further from the nucleus than the 2s-like states, so they will see more screening of the positive charge of the nucleus by the 1s-like electrons, so they will be slightly less tightly bound than the 2s-like states.

As we keep adding electrons, we fill the 2s-like and 2p-like states, completing another shell. Since there are three different p orbitals, and we are positing two spin states for each, it will take a total of 8 electrons to fill this “shell” – 2 to fill the two 2s-like states (one for each spin), and 6 to fill the six 2p-like states (3 for each spin). On this simple model, the next “shell” would have two 3s electrons, six 3p electrons, and ten 3d electrons (there are 5 d orbitals, corresponding to $m = -2, -1, 0, 1, 2$), for a total of 18. As we fill the various “shells”, the additional electrons are seeing progressively weaker attraction to the nucleus because its positive charge is being partially screened by more and more electrons.

We can imagine continuing at least approximately with a process like this. So, generally, in these many-electron atoms, we expect orbitals of higher n to be progressively less tightly bound¹⁴. We also expect orbitals of higher l for a given n to be less tightly bound, because they are generally somewhat farther from the nucleus and hence will see a nucleus that is more screened by the closer electrons. As we noted, the “inner” orbitals with smaller n will be very strongly bound, even more so that in the hydrogen atom, because they are seeing a larger net positive charge on the nucleus. Hence in these many electron atoms, the “core” electrons of smaller n are very tightly bound; as a result, they mostly do not participate in chemical reactions, which instead primarily involve the relatively weakly bound “outer” electrons at higher n and l .

In general, then, we expect the orbitals to “fill up” in some way in which the states of lower n are filled first, because they are more tightly bound, and to some extent also we expect the states of lower l to have lower energy than those of higher l and so to be filled earlier as well. There is an empirical rule, known as Madelung’s rule, which is more sophisticated than our simple “shell” approach above. It is approximately (but not perfectly¹⁵) followed by most atoms, and mostly gives the correct order of the filling of the orbitals.

Madelung’s rule states that

¹⁴ Of course, in the hydrogen atom, states of higher n are already less tightly bound, but the decrease in binding strength with increasing n is magnified further here by the increased “screening” of the charge of the nucleus by the more tightly bound electrons in the inner shells.

¹⁵ See T. L. Meek and L. C. Allen, “Configuration irregularities: deviations from the Madelung rule and inversion of orbital energy levels,” *Chem. Phys. Lett.* 362, 362-364 (2002) doi:10.1016/S0009-2614(02)00919-3 for a list of exceptions. Common elements that do not obey the rule include chromium, copper, silver, platinum, and gold.

the orbitals are filled up starting from smaller values of $n + l$ and proceeding to larger values of $n + l$ with the states of lower n being filled first for each particular value of the sum $n + l$.

This approximate rule is an example of the general principle¹⁶ that the states will be filled up in order of their energy, starting from the lowest (most tightly bound) states and proceeding to higher energy states.

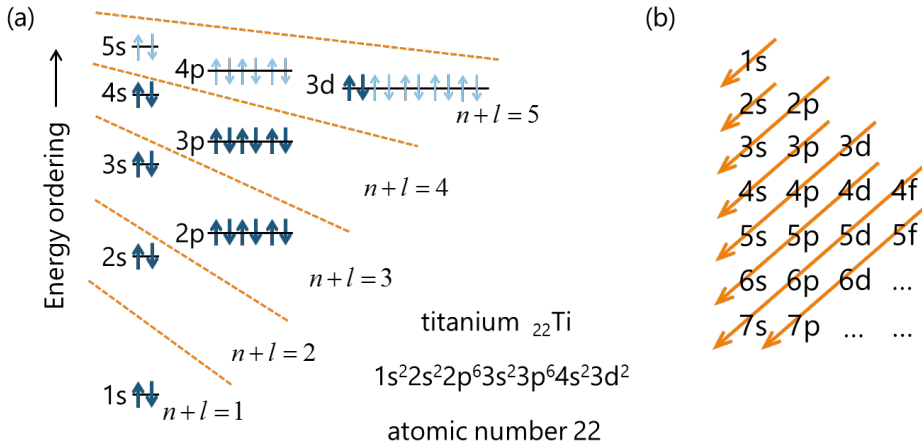


Fig. 4.7 (a) Ordering of the energies of orbitals according to Madelung's approximate rule, grouped by the sum $n + l$. The short arrows indicate electrons of spin "up" and spin "down". An example is shown for the titanium atom, in which 22 electrons fill the orbitals according to this rule. The energy is not to scale, and indicates only the ordering of the orbital energies, not actual energies. Light colored short arrows indicate unoccupied orbitals. (b) A simplified diagram indicating the order of filling the orbitals according to Madelung's rule. The filling proceeds first down the highest arrow, in this case just the $1s$ orbital, then to the next arrow, so the $2s$ orbital, then to the next arrow, so the $2p$ orbitals first, then the $3s$ orbitals, and so on.

Madelung's rule is illustrated in Fig. 4.7. An explicit example is also given for the titanium atom, which has 22 electrons. In this case, all the orbitals up to an including $n + l = 4$ are fully occupied, and two electrons are in the $3d$ orbitals. This figure also illustrates some common notations. The chemical symbol for titanium is Ti. The number of protons in the titanium nucleus, the atomic number, is 22, which is also the number of electrons in the neutral titanium atom (that is, when it is not ionized in any way). The atomic number can also be indicated explicitly together with the chemical symbol as the subscript 22, as in ${}_{22}\text{Ti}$. The number of electrons in each of the sets of orbitals is indicated using the superscripts in the atomic configuration, here $1s^2 2s^2 2p^6 3s^2 3p^6 4s^2 3d^2$.

In writing out such configurations, there is a commonly used "shorthand" way of writing them, based on the configurations of the various noble gasses. So we can write

¹⁶ This is known as the Aufbau principle.

$$\begin{aligned}
[\text{He}] &\equiv 1s^2 \\
[\text{Ne}] &\equiv 1s^2 2s^2 2p^6 \equiv [\text{He}] 2s^2 2p^6 \\
[\text{Ar}] &\equiv 1s^2 2s^2 2p^6 3s^2 3p^6 \equiv [\text{Ne}] 3s^2 3p^6 \\
[\text{Kr}] &\equiv 1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 \equiv [\text{Ar}] 3d^{10} 4s^2 4p^6 \\
[\text{Xe}] &\equiv 1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 5s^2 5p^6 \equiv [\text{Kr}] 4d^{10} 5s^2 5p^6
\end{aligned} \tag{4.27}$$

With this shorthand, for example, the titanium configuration can be written

$$1s^2 2s^2 2p^6 3s^2 3p^6 4s^2 3d^2 \equiv [\text{Ar}] 4s^2 3d^2 \tag{4.28}$$

Note that the order of writing the configuration does not matter. The above version, (4.28), for titanium is written in a “Madelung’s rule” order. It is also common to write configurations first in order of increasing n , and then in order of increasing l within each group with the same n , as in

$$1s^2 2s^2 2p^6 3s^2 3p^6 3d^2 4s^2 \equiv [\text{Ar}] 3d^2 4s^2 \tag{4.29}$$

The above configurations, (4.27), for the noble gasses have been written this way.

This kind of approach to many-electron atoms largely explains the electronic configuration of atoms, and also explains why it is the electrons in the last “shell”¹⁷ that participate in most of the chemical reactions, those being the least tightly bound electrons in the atom.

Quiz questions

Q4.8.1 For each of the following proposed atomic configurations, state whether they obey Madelung’s rule or not.

- a) Carbon (atomic number =6): $1s^2 2s^2 2p^2$
- b) Neon (atomic number =10): $1s^2 2s^2 2p^6$
- c) Aluminum (atomic number =13): $1s^2 2s^2 2p^6 3s^2 3p^1$
- d) Chromium (atomic number =24): $1s^2 2s^2 2p^6 3s^2 3p^6 4s^1 3d^5 \equiv [\text{Ar}] 4s^1 3d^5$

Problems

4.8.1* Suppose I have an atom in which all of the 1s, 2s, 2p, and 3s orbitals are occupied, and the total number of electrons in the atom is equal to the atomic number of the atom. What is the atomic number of the atom? What element is the atom?

4.8.2* The Madelung rule states that the orbitals are populated by electrons in the increasing order of the sum of quantum numbers $n + l$, with priority to the orbitals with smaller n .

¹⁷ The electrons in the last partially occupied shell are known as the “valence” electrons in chemistry.

Applying this rule to the potassium atom (which has atomic number 19 and the symbol K, and can be written using a standard notation as ${}_{19}\text{K}$), we see that its $n = 4, l = 0$ set of orbitals is partially occupied, while the $n = 3, l = 2$ orbitals are empty. This is the first atom in the periodic table of the elements to fill a shell with a higher n before exhausting all orbitals with lower n (though it is still obeying the Madelung rule). Write out the quantum numbers n, l of the occupied orbitals in the ground state of ${}_{19}\text{K}$ and point out how many electrons are associated with each.

- 4.8.3 Presuming the Madelung rule works, what is the minimum number of electrons needed to occupy orbitals in such a way that an orbital with quantum number $n = q$ (for some integer q) is filled before all orbitals with $n = q - 2$ are completely filled? Which atom is the first one to exhibit this property? (Expressed in similar terms, the potassium atom (as in the preceding problem) is the first one, for example, for which an orbital with $n = j$ is filled before all orbitals with $n = j - 1$ are completely filled, with, in that case, $j = 4$.) (Once you get your answer to the number of electrons, you may need to look at the periodic table of the elements. Incidentally, the resulting atom does obey Madelung's rule.)
- 4.8.4 For each electron configuration given below, all of which are believed to be the electron configurations for actual atoms, state whether it corresponds to Madelung's rule, explaining your answer briefly if you believe it does not correspond to Madelung's rule.

In each case, also state what element is being represented by this configuration [Hint: count the total number of electrons, and hence deduce the element from a periodic table or otherwise].

- (i) $1s^2 2s^2 2p^6 3s^1 \equiv [\text{Ne}] 3s^1$
- (ii) $1s^2 2s^2 2p^6 3s^2 3p^1 \equiv [\text{Ne}] 3s^2 3p^1$
- (iii) $1s^2 2s^2 2p^6 3s^2 3p^6 3d^5 4s^1 \equiv [\text{Ar}] 3d^5 4s^1$
- (iv) $1s^2 2s^2 2p^6 3s^2 3p^6 3d^6 4s^2 \equiv [\text{Ar}] 3d^6 4s^2$
- (v) $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^1 \equiv [\text{Ar}] 3d^{10} 4s^1$
- (vi) $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^6 5d^4 6s^2 \equiv [\text{Xe}] 4f^{14} 5d^4 6s^2$
- (vii) $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^6 5d^{10} 6s^1 \equiv [\text{Xe}] 4f^{14} 5d^{10} 6s^1$

4.9 Fermions and bosons

Spin and elementary particles

The idea of spin, originally introduced empirically to construct rules for explaining configurations of electrons in atoms, has much deeper and broader uses in explaining the physics of the world around us.

Spin is an intrinsic property of all “elementary” particles, and its magnitude is fixed for each type of particle. Particles divide into two categories on the basis of their spin:

fermions¹⁸, which have half-integer spin

bosons¹⁹, which have integer spin

Of the various particles we come across in daily life

electrons, protons, and neutrons all have spin $\frac{1}{2}$, and are fermions

photons have spin 1, and are bosons

All of the “elementary particles”, that is, ones that are not apparently made up out of other particles, either appear to have spin $\frac{1}{2}$ or spin 1, with the exception of the Higgs boson, which is thought to have spin 0, and the graviton which is posited to have spin 2.

It is possible to make “compound” particles. The proton and the neutron are themselves thought to be compound particles, each being made up out of three “quarks”. Quarks themselves are fermions. So, strictly, protons and neutrons are not elementary particles, though from a practical point of view in the everyday world we treat them as if they are.

We can have other compound particles with integer or half-integer values of spin, with those values being made up out of sums and/or differences of the spins of the constituent (more) elementary particles. So, for example, the helium-4 nucleus (consisting of two protons and two neutrons, also known as an “alpha particle” in radioactivity) behaves like a boson, as does the helium atom, which has two electrons together with a helium-4 nucleus²⁰.

Perhaps the single most important difference between fermions and bosons is that fermions obey the Pauli exclusion principle but bosons do not. So, we can have any number of bosons all in the same state.

One particularly important practical consequence of the fact that bosons do not obey the Pauli exclusion principle is the laser. In the laser, we can have very large numbers of photons (which are bosons) all in the same “mode”, which in practice means all in the same light beam.

One common misconception is to think that electrons interact because they are fermions and photons do not interact because they are bosons. Electrons interact because they are charged. Photons have extremely weak interactions²¹ because they are not charged. Helium-4 nuclei, which are bosons, would interact very strongly, of course. Part of the confusion is that we say that we can have multiple bosons in the same state. For strongly interacting bosons like helium-4 nuclei, that is indeed still true, but an eigenstate occupied by some large number of helium-4 nuclei would not necessarily be the same

¹⁸ Named after the Italian physicist Enrico Fermi.

¹⁹ Named after the Indian physicist Satyendra Nath Bose

²⁰ Helium can also exist in the helium-3 form, with a nucleus that only has one neutron together with the two protons. Though relatively rare compared to helium-4, it is thought to be stable – that is, not radioactive. Helium-3 atoms behave as fermions.

²¹ It is possible for two photons to interact in the vacuum to produce an electron-positron pair, for example.

eigenstate that we would put one helium-4 nucleus in; the interaction between the bosons would in general change the eigenstates of the system.

Identical particles

Another subtle point about these quantum mechanical particles is that particles of a given type are absolutely identical to one another, with no way whatsoever of telling them apart. This means that the counting of possible states of multiple particles is different from that of classical particles.

One way of understanding this difference is to compare dollar bills and dollars in bank accounts. Dollars in bank accounts are identical in the sense that quantum mechanical particles of a given kind are identical. If we have two dollar bills and two boxes, there are two different ways of having one dollar bill in each box – bill 1 in box A and bill 2 in box B, or bill 2 in box A and bill 1 in box B. But if we think of two bank accounts, there is only one way of having one dollar in bank account A and one in bank account B. This “identity” of quantum mechanical particles changes the way that they distribute themselves among available states when we think about thermal distributions.²²

Quiz questions

Q4.9.1 State whether or not each of the following particles obeys Pauli exclusion

- a) Electron
- b) Proton
- c) Photon
- d) Neutron
- e) Helium-4 nucleus

Q4.9.2 Suppose I have three marbles-- one red, one green, and one blue-- and two containers, a vase and a shoebox. How many different ways are there to put the three marbles in the two containers?

Q4.9.3 How many ways are there of having 3 dollars into two different bank accounts (e.g., one checking, one savings) (without splitting the dollars up into sub-dollar amounts)?

²² We generally think of classical particles as being different from one another in the way that, say, bricks are similar but slightly different from one another. For such classical particles that are similar but not identical to other such particles, when the particles on the average have some thermal energy, those classical particles will distribute themselves among the various possible states according to the Maxwell-Boltzmann distribution. For identical fermions, they will obey the Fermi-Dirac distribution, and for identical bosons, they will obey the Bose-Einstein distribution.

Problems

- 4.9.1 (i) Suppose you have 3 dollar bills (all with different serial numbers, of course!). How many different ways are there of distributing these dollar bills between two boxes A and B (including the possibility of having no dollars in a given box)?
- (ii) Suppose you have 3 dollars and two bank accounts, one with bank A and one with bank B. How many different ways are there of distributing the 3 dollars (allowing only integer numbers of dollars) between the two bank accounts (including the possibility of having no dollars in a given bank account)?
- 4.9.2 For each of the following particles, state whether it should obey the Pauli exclusion principle.
- (i) An electron
 - (ii) A photon
 - (iii) A neutron
 - (iv) A Higgs boson
-

4.10 Coupled systems

So far in quantum mechanics, we have considered mostly single quantum mechanical systems, like a particle in a box or the hydrogen atom. Though these are useful models and abstractions, much of quantum mechanics concerns how systems interact with one another. Multiple atoms can form molecules or crystalline structures, for example.

The details of such systems can become complicated, but we can start to understand some of the key features of such systems by looking first at one of the simplest – two potential wells separated by a barrier of finite height and width. Such a structure is often called a “coupled well”. Such a system is shown in Fig. 4.8. For each of the cases shown in Fig. 4.8, we have calculated the first two energy eigenvalues (or energy “levels”) and the associated eigen functions (wavefunctions) in each structure, including the effects of the finite heights of the potential barriers on either side²³.

At the bottom of Fig. 4.8, we see one simple limiting case, in which there is no barrier (the barrier is of “zero thickness”). Here we see two solutions that are similar to the

²³ We could perform such calculations using the methods we have been developing, including the boundary conditions and appropriate forward and backward waves in each region. The case of zero barrier thickness corresponds to a “finite potential well”, which can be solved relatively straightforwardly, though the solutions for the eigen energies are not in closed form – we have an expression to calculate them, but this is not a simple formula, requiring us to find zeros of a function numerically. For the more complicated coupled-well problem, we can in principle take a similar approach. In practice we use other matrix methods to help us handle the large number of simultaneous equations. The calculations here for the coupled wells were performed using one such matrix approach, the “transfer matrix” method. See, e.g., *Quantum Mechanics for Scientists and Engineers* by D. A. B. Miller (Cambridge, 2008), Chapter 11.

solutions of the potential well with infinitely high barriers. Because the barriers here are of finite height, there is some tunneling penetration into the barriers as we can see.

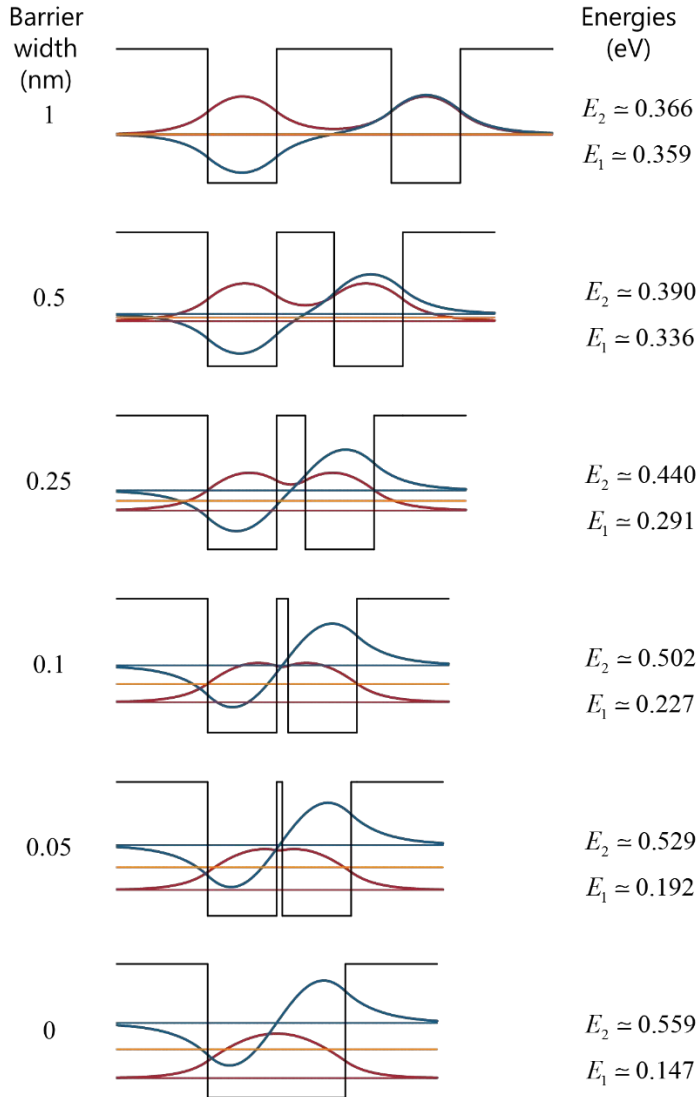


Fig. 4.8. The first two energy levels and eigenfunctions of a coupled potential well for an electron, shown for various values of the width or thickness of the barrier between the wells. The height of the potential energy on both sides and in the barrier is 1 eV relative to the bottom of the wells. Each of the two wells is 0.6 nm thick. The orange horizontal line (generally, the one in the middle of the set of horizontal lines) shows the energy of the first level (at ~ 0.362 eV) in an isolated potential well of 0.6 nm width, with the same 1 eV barrier height. The red and blue horizontal lines show the energies of the corresponding first (lower energy) and second (higher energy) states of the coupled well for each barrier width, and these red and blue lines are used also for the vertical origins for plotting the corresponding eigenfunctions.

The lowest solution is symmetric about the center of the well, and the second solution is antisymmetric (opposite signs on opposite sides of the center), just as in the box with infinitely high walls. As we increase the thickness of the barrier, we retain these symmetry properties. The first, lower-energy solution in particular, however is changing its character, acquiring a “dip” in the middle. The second solution initially does not change much in character, though as the barrier gets thick, the central portion of this wave function – the part in the barrier – become “stretched”.

For the thickest barrier shown in Fig. 4.8 (1 nm), the solutions as seen within a given well are essentially similar to the solution for a single isolated well with the same width and barrier height. Note, however, that the solutions of this coupled-well problem are not the same as those of two isolated wells; these are always coupled solutions; each solution describes a state of the coupled pair of wells.

In the limit of a thick barrier, we end up with two eigenfunction solutions: one solution is essentially the “symmetric” sum of the individual “isolated-well” wavefunctions – that is, in this case, a single “bump” in the right well plus a single “bump” in the left well; the other solution is essentially the “antisymmetric” sum of those “isolated-well” wavefunctions – that is, in this case, a single “bump” in the right well minus a single “bump” in the left well²⁴. For a thick barrier, the energies of these solutions are approximately equally spaced below (for the “symmetric” solutions) and above (for the “antisymmetric” solutions) the energy of the first state of an isolated well²⁵.

One might think that, especially as we consider thick barriers, those “isolated well” solutions would become valid again. In fact, though, they are not energy eigenfunctions of this coupled problem, and they would not be stable in time. This situation is like that of two coupled “masses-on-springs” problems, or of two pendulums coupled by a weak spring. If we excite one of the masses oscillating, it will gradually transfer its oscillation over to the other one; subsequently, that oscillation will transfer back over again to the first mass, and so on. If we “started” the electron in only one of the wells, the same oscillatory behavior would result here also, with the electron oscillating back and forward between the two wells²⁶. We could think of this, loosely, as the electron tunneling from one well to another and then tunneling back again, and so on.

²⁴ Whether this antisymmetric solution is “right” minus “left” or “left” minus “right” is of no particular significance or meaning. Indeed, quite generally, we can multiply any such eigenfunctions by any unit complex factor (e.g., $\exp(i\theta)$) and they will still be valid eigenfunctions (as long as we multiply the entire eigenfunction by this same constant factor). What matters for the antisymmetric solution is that what is on the “left” is minus the value of its mirror image point on the “right”.

²⁵ One simple approximate model for the case of thick barriers is the so-called “tight-binding” model. In that tight-binding model, which does predict the limiting behavior for thick barriers, the two levels are indeed split to be equally below and above the “isolated well” level, and the wavefunctions are these symmetric and antisymmetric sums, respectively, of the two different “isolated well” wavefunctions.

²⁶ In quantum mechanics, when we extend to the time-dependent Schrödinger equation, we find an angular frequency ω associated with an energy E through $E = \hbar\omega$. The electron would oscillate back and forward between the wells with a frequency $\omega_{21} = (E_2 - E_1) / \hbar$.

In both the classical coupled oscillators (or masses on springs) and our case here, for the eigen solutions for this symmetric problem (identical masses and springs for the classical problem²⁷ and identical wells for our quantum problem) this kind of symmetric and antisymmetric behavior for the eigen states is retained even as the coupling becomes strong.

Another very important point about this coupled solution is that the coupling here *reduces* the energy of the lower state compared to the energy of the isolated well. This phenomenon is at the core of the idea of covalent bonding of atoms in molecules. It is the reason why two identical atoms can bond with one another – their energy is reduced in one of the coupled states. Of course, there are many more details in the actual analysis of such bonding, but this notion – that sharing of an electron between two potentials can reduce the energy of the state – is very important. We see a growth of the probability density of finding the electron in the region “between” the two wells; generally the electron is on the average somewhat nearer to the center of the coupled system in this state.

We can see from Fig. 4.8 why the energy of the lower coupled state is reduced compared to the isolated well state; this is particularly clear as we look at the cases of strong coupling (thinner barriers). For example, for a barrier thickness of 0.25 nm, we can see quite clearly that the lower energy wavefunction is less strongly curved overall within a given well than the “isolated well” wavefunctions (which are very similar to the wavefunction within a given well in the 1 nm barrier case). That lower curvature corresponds to lower kinetic energy for the electron in the Schrödinger wave equation (a smaller magnitude of the second derivative of the wavefunction). Symmetric wavefunction states like this are sometimes called “bonding” states.

The second, “antisymmetric” state corresponds to higher energy than the isolated-well state. This wavefunction requires larger curvature because it must go through zero in the barrier. That larger curvature corresponds to larger kinetic energy (a larger magnitude of the second derivative of the wavefunction). Because of its higher energy, this kind of antisymmetric state is sometimes called an “antibonding” state.

Quiz questions

Q4.10.1 Suppose we have two identical “rectangular” potential wells that we are going to couple by bringing them closer together to make a combined system. Considered separately, each such well would have a quantum mechanical energy eigenstate for an electron, with some energy E_{01} relative to the bottom of the potential well. For the purposes of this question, we presume that each such well would only have this one possible electron energy eigenstate. As we bring these wells closer together in the combined system, consider the following statements, and choose the one that is correct.

- a) The combined system has two possible energy eigenstates for an electron, and both these states have the same eigenenergy

²⁷ The spring providing the coupling between the classical masses need not be identical to the other two springs.

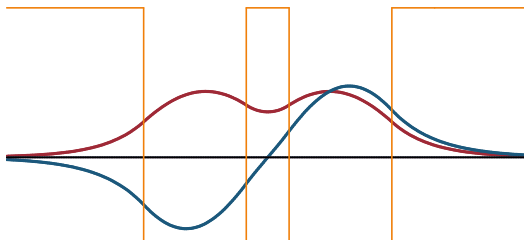
- b) The combined system has two possible energy eigenstates for an electron, and these states each have eigenenergies that are different from E_{01} .
- c) The combined system has four possible energy eigenstates for an electron, two with eigenenergy E_{01} , and another two with different eigenenergies.

Q4.10.2 For the same system as in Q4.10.1 above, choose whether the following statements are true or false.

- a) As we compare the lowest two energy eigenstates, relatively speaking we more likely to find the electron in the barrier region when the electron is in the lower energy state than when the electron is in the upper energy state.
- b) There is an energy eigenfunction solution that is symmetric with respect to the center of the barrier.
- c) As the barrier width shrinks, each energy eigenfunction solution changes from being symmetric to antisymmetric (with respect to the center of the barrier).
- d) The “bonding” state has higher energy because it corresponds to a more tightly bound system.

Problems

4.10.1 Consider a coupled potential well with potential barrier heights $V_0 = 1$ eV, well thickness $(L_z) = 0.6$ nm and barrier thickness = 0.25 nm. The first two eigenfunctions (ψ_1 and ψ_2) are shown below. These wavefunctions are plotted on the same scale below, and relative to the same vertical zero, which is given by the horizontal line. Both of these wavefunctions are normalized on the same vertical scale, so the relative amplitudes of the two wavefunctions are meaningful. The “first” eigenfunction ψ_1 is positive everywhere on this graph, and the “second” eigenfunction ψ_2 is positive and negative. The form of the potential is also sketched on the same horizontal scale.)



- (a) (i) In which of the two eigenstates is an electron more likely to be found in the middle barrier region ($-0.125 \text{ nm} < z < 0.125 \text{ nm}$)?
- (ii) In which of the two eigenstates is an electron more likely to be found in the outer barrier region ($|z| > 0.725 \text{ nm}$)?
- (b) We will say that an electron in the first eigenstate (the lower energy one) has the probability of P_1 to be found in the middle barrier region and an electron in second eigenstate has the probability of P_2 to be found in the

middle barrier region. If the middle barrier thickness reduces from 0.25 nm to 0.05 nm, would you expect the ratio between P_1 and P_2 to increase or decrease? Why?

- 4.10.2 Consider the calculated energy levels E_1 and E_2 as given in Fig. 4.8 for each of the different thicknesses d of the barrier, from 0 to 1 nm, and specifically the energy separation between the two, $\Delta E = E_2 - E_1$. By intelligent guesswork and empirical comparison, deduce a simple expression for ΔE that fits the calculated values well (e.g., to better than a few meV).

[Hints: Graph up the values of ΔE as a function of the barrier width, look at the results, guess a form (you may even have some physical intuition here as to what the form “ought” to be), plot that form on the same graph and adjust the parameters empirically until you get a good fit. This is likely easiest done with some kind of computer program or software. You may also want to create a table of the differences between the given values of ΔE and your model values. You could use that table when fine-tuning your parameters. You should not need to use any formal “curve fitting” to get your result, and you should be able to come up with a very good model with only two parameters, one of which can be chosen very simply and obviously.]

4.11 Crystals

The combination of atoms into larger structures is, of course, very important in a wide range of applications. A very large part of chemistry is based on such combinations of atoms into molecules. One particularly important category of structures for applications in engineering is crystals, which are essentially large, regular structures of many atoms. A large fraction of the devices we use to process and communicate information, such as diodes, transistors, photodetectors, and semiconductor lasers, and for other applications, such as solar cells, use so-called “semiconductor” materials. Most of these semiconductor materials are crystals. Here we will introduce some of the key features and processes in these crystal structures that we will need for a basic understanding of how many such devices work.

Many simple materials will form themselves in crystalline structures of atoms. Crystals have several advantages as materials. If we are careful about the environment and manner in which we “grow” crystals, they come out essentially exactly the same every time, so they are very predictable materials. The regularity of their structure means that, even though there may be a vast number of atoms in a piece of crystal, we can have relatively simple models by which we can understand their properties, such as their electronic or optical behavior.

With semiconductors in particular, it is also often the case in practice that introducing small amounts of controlled impurities or “dopants” in the structure enables us to get some very precise and important kinds of control over the detailed electronic properties of the material.

In a mathematical sense,

a crystal is a structure that is made up by stacking identical “bricks” (or “unit cells”) together in a regular fashion in such a way that they fill up all space.

(4.30)

Essentially any repeating pattern, such as a wall paper pattern, is crystalline in this way. A simple two-dimensional lattice is shown in Fig. 4.9. We can think of this as a collection of rectangular “bricks”, or in the terminology of crystals, rectangular “unit cells”. As we can see, stacked together, these unit cells will fill all space (in two dimensions).

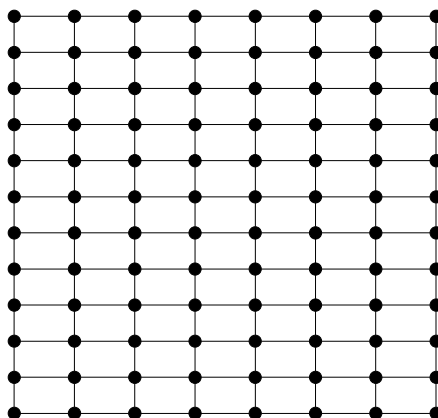


Fig. 4.9. Example of part of a two-dimensional rectangular crystal lattice

We can also view the crystal as a lattice of points (the “black dots” in Fig. 4.9). We can always make such a lattice by marking the same position on every brick or unit cell with such a “black dot”, or, more correctly, a mathematical point. It does not matter which point we choose in the unit cell for this. Explicitly, this set of mathematical points forms a crystal lattice. Note that the lattice is not quite the same as the crystal; the lattice is just this set of points, one per unit cell, whereas the crystal is the entire structure made using the unit cells.

Often when we draw a lattice, we will connect the points by lines to guide the eye, as in Fig. 4.9, though these lines are not essential mathematically. Once we start thinking about a crystal being made up of atoms that are chemically bonded to one another, we may also like to draw such lines, with the lines often representing, in a loose way, the chemical bonds between the atoms. However, these lines may not always have any direct relation to actual chemical bonds.

Three-dimensional crystals are slightly more difficult to visualize than two-dimensional ones, but they follow the same rules. In semiconductor electronic and optoelectronic devices, many of them are based on one of the closely related lattices shown in Fig. 4.10. The diamond lattice is the crystal lattice for the semiconductor materials silicon and germanium. The “zinc blende”²⁸ lattice is the lattice for the many of the so-called

²⁸ Zinc blende is zinc sulfide in a cubic crystalline form.

“compound” semiconductor materials based on “III-V” materials.²⁹ III-V compound semiconductors are materials made from a combination of elements from Group III of the periodic table (e.g., gallium, aluminum, indium) and Group V of the periodic table (e.g., arsenic, phosphorus, antimony, nitrogen). Typical III-V semiconductors include gallium arsenide and indium phosphide.

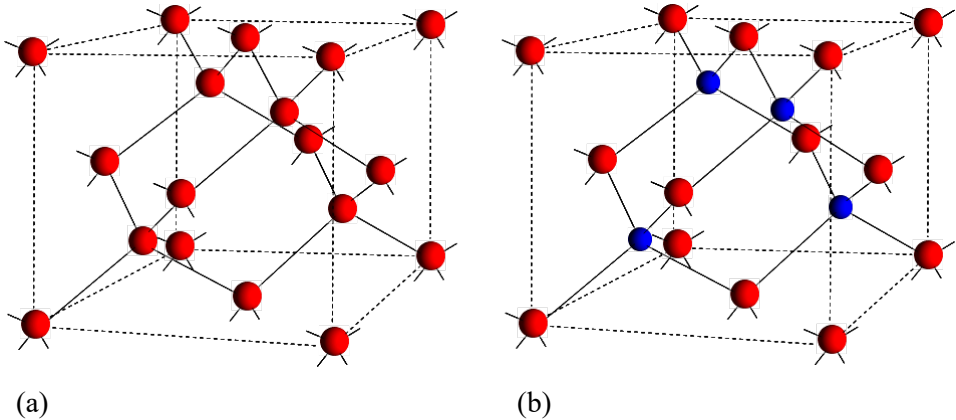


Fig. 4.10. (a) Diamond lattice. (b) “Zinc blende” lattice.

Both the diamond and zinc blende lattices are based on atoms forming four chemical bonds, equally spaced in angle, with their neighbors. Silicon and germanium are both Group IV elements, which means four of their outer eight electron orbitals are filled. This fact makes them particularly want to bond with four other atoms. Carbon is also a Group IV element, and when it crystallizes in this form, it is called diamond (hence the name of the lattice structure).³⁰ In these lattices, the four atoms connected to a given atom form the corners of what is called a (regular) tetragon – if we joined these four atoms with lines (which in this case would in a sense correspond to chemical bonds), they form a structure with four equilateral triangle sides, all of equal size.

A close look at the diamond lattice shows that, somewhat surprisingly, a cubic structure emerges. The cube is indicated by the dashed lines (and in this case those dashed lines are *not* representing chemical bonds). If we look carefully, we see that there is an atom on each corner of the cube, and one in the middle of each face. A structure consisting of sets of atoms arranged like that is called a “face-centered cubic” lattice.

²⁹ One exception is gallium nitride, and some of its alloys, which can crystallize in a so-called “wurtzite” structure. The wurtzite lattice is like the zinc blende lattice, but in one direction the alternate planes are rotated by 60 degrees compared to those of the zinc blende lattice. Gallium nitride is useful in short wavelength (e.g., blue) light-emitting diodes, for example.

³⁰ Carbon can also crystallize in another completely different form, graphite, which is most common as the “lead” in a pencil. The mechanical properties of graphite are completely different from diamond; graphite is quite soft, and is a good lubricant, whereas diamond is extremely hard, and is a good abrasive! Graphite is made of loosely bound layers of carbon atoms. These layers can be detached relatively easily, in which case the layers are called graphene. If we “roll up” graphene sheets, we can make so-called carbon nanotubes. Related structures can also be formed into more spherical structures called “bucky balls”.

There are also other atoms that do not lie on the corners or the faces. If we were to continue this lattice by adding more atoms in their crystalline positions in adjacent unit cells, we would see that these other atoms also lie on a face-centered cubic lattice that is simply displaced by one chemical bond in space. Thus the diamond lattice in this sense is actually made from two interlocking face-centered cubic lattices (or sometimes described as interlocking sublattices).

In the zinc blende form of lattice, one of the kinds of atoms (e.g., arsenic, represented by the larger balls in Fig. 4.10) lies on one of these face-centered cubic sublattices, and the other (e.g., gallium, represented by the smaller balls in Fig. 4.10) lies on the other face-centered cubic sublattice. Though the Group III element only has 3 electrons in its outer shell, and the Group V has 5, they share these between the atoms to make four bonds for each atom. The fact that, in the zinc blende lattice, one kind of atom sits on one of the face-centered cubic sublattices and another kind lies on the other face-centered cubic sublattice helps us clarify that we indeed have two interlocking face-centered cubic lattices in both the zinc blende and diamond lattices³¹.

An important additional set of materials, especially for optoelectronic devices, is the set of III-V alloy materials. These are materials such as indium gallium arsenide, or aluminum gallium arsenide. Materials with three elements in them are known as “ternary” alloys. In these example cases, the Group III face-centered cubic sublattice has a usually random distribution of, e.g., indium and gallium atoms on it. It is also possible to extend this to choices of two atoms for each of the sublattices, as in indium gallium arsenide phosphide (which, because it has four components, is called a “quaternary” alloy), or three different choices for one lattice, as in indium gallium aluminum arsenide (also a “quaternary” alloy). Though the random distribution of elements on a given sublattice means these alloys are technically not crystals, in practice we can mostly get away with treating them as crystals with some kind of average of the properties of the perfect two-component “binary compound” crystals (such as gallium arsenide and indium arsenide). These alloys give a great deal of flexibility in the design of materials with exactly the desired properties, and they are used extensively in

³¹ Now, we might ask, for both the diamond and zinc blende cases, what is the mathematical crystal lattice that corresponds to these? The answer to that question is that the crystal lattice in the mathematical sense is actually one face-centered cubic lattice. But, for each such mathematical point in the lattice, there are *two* adjacent atoms associated with that point. For the zinc blende case, that pair of atoms corresponds to one of each kind (e.g., one Group III atom and one Group V atom). So in that sense the zinc blende lattice is actually a lattice of *pairs* of atoms (one atom from each “sublattice”). The actual unit cell for such a crystal will generally have more than one atom in it. There are often many ways we can draw a unit cell for a given lattice, all of which may be valid (that is, stacking them regularly side by side fills all space), and different unit cells may have different numbers of atoms. So-called “primitive” unit cells are ones with one lattice point in them, and one well-known class of such primitive unit cells is the Wigner-Seitz unit cell, which can be formed for any lattice. For this zinc blende crystal, however, even a primitive unit cell will have at least two atoms in it (one of each kind). This emphasizes for us again that we should not necessarily associate each lattice point with an atom, nor should we necessarily associate the lines between the crystal lattice points with chemical bonds.

optoelectronic devices. Changing the alloy composition can change the wavelength or color of light emitted by a light-emitting diode or semiconductor laser, for example.

Of course, not all the materials we use in making devices are crystalline. Often, we take care to make the semiconductors crystalline because we need the best, most predictable, and precisely controlled performance from them. Other materials also used in making devices, such as metals for conductors and oxides for insulators, do not have to have precise crystalline forms to operate successfully, and we can be more relaxed in the way we make them³². Materials where we are not too concerned about a precise crystalline structure can often be made by simple, flexible techniques, such as evaporation onto the surface of interest. Growth of crystals usually requires carefully controlled temperatures; also, crystals can typically only be grown uniformly and precisely if they are being grown on some substrate or seed that itself is crystalline with the same separation between the atoms; such growth is called “epitaxial” growth, meaning that the growing surface takes on the order of what lies beneath it³³. Materials with essentially no crystalline structure are called “amorphous”; amorphous silicon is relatively easy to make and can be used in solar cells, for example.

Quiz questions

Q4.11.1 We know we can pack spheres tightly together in regular structure, like stacking oranges for example. For the purposes of this question, we can presume this packing goes on forever in space (i.e., the stack of oranges is infinitely large in all directions). The distance between the centers of two spheres that are touching in this stack is, of course, the diameter d of the spheres. Choose whether each of the following statements is true or false.

- a) The centers of the spheres form a crystal lattice
- b) The spheres are unit cells for a crystal structure
- c) Suppose you closed your eyes and I moved the stack of spheres by an amount d along the direction between the centers of two touching spheres, then when you opened your eyes, you would not be able notice any difference.

Q4.11.2 In the diamond crystal structure, choose whether each of the following statements are True or False.

- a) Each of the atoms lies on the corners of set of non-overlapping cubes that are stacked to fill all space
- b) We could use a cube as a unit cell for this crystal structure.

³² Elemental metals can also be made as perfect crystalline materials, and this is sometimes done for specific purposes, including scientific studies. Silicon oxides, used extensively as insulators in electronic integrated circuits, are usually not crystalline, though a crystalline form of silicon dioxide, known as crystalline quartz, can be made and is used for various applications. Glass, which is also essentially silicon dioxide, is actually best classed as a super-cooled liquid, and is quite amorphous.

³³ In “epitaxial”, “epi” means “surface” (as in “epidermis” for the outer layer of skin), and the “tax...” part means an “ordering” (as in “taxonomy” for classification).

- Q4.11.3 If we have one atom for each crystal lattice point, all the atoms in the lattice have to be the same kind of atom. True or False?
- Q4.11.4 There is more than one form of crystal structure for carbon atoms. True or False?
- Q4.11.5 A binary crystal is made up of atoms of two and only two different kinds. True or False?
- Q4.11.6 A ternary alloy is made up of four different kinds of atoms
- Q4.11.7 All the materials used in making integrated circuits are in crystalline form. True or False?

Problems

- 4.11.1 If the following forms of matter were infinite in size, would they be considered as crystals and why or why not? (Note: To avoid any confusion, use the specific mathematical definition of a crystal given above in (4.30).) (You may want to consult other reference materials if you need to understand more about the various forms of matter below.)
- (a) Liquid water
 - (b) An ice cube (at least, an ideal one)
 - (c) Water vapor
 - (d) Window glass
 - (e) A snowflake
 - (f) Diamond
 - (g) Graphite
 - (h) A silicon wafer
 - (i) Wax
 - (j) Graphene

4.12 Emergence of bands

We discussed above that electrons in atoms have specific energy levels they can occupy. For every atom of a given type, those levels are, of course, the same. But what happens to the energy levels when we start to bring two or more atoms closer together, as must happen as we form a solid?

It is quite general that, when we bring two identical systems together, and start to allow them to interact, the energy levels split, as we saw in the example of the coupled well above. If we increase the number of identical systems we couple, we correspondingly increase the number of levels into which a given “atomic” or isolated-well level splits. In fact, for N identical coupled systems, each of the original levels will tend to split into up to N distinct ones. This is illustrated in Fig. 4.11 for the cases of two, four, and eight

coupled wells³⁴. The detailed wavefunction solutions for the eight well case are shown in Fig. 4.12. The case shown in Fig. 4.11 and Fig. 4.12 is for a relatively narrow well so that we can clearly see the splitting of the first level. The parameters for an individual well in Fig. 4.11 and Fig. 4.12 are the same as in the upper panels of the coupled wells of Fig. 4.8³⁵.

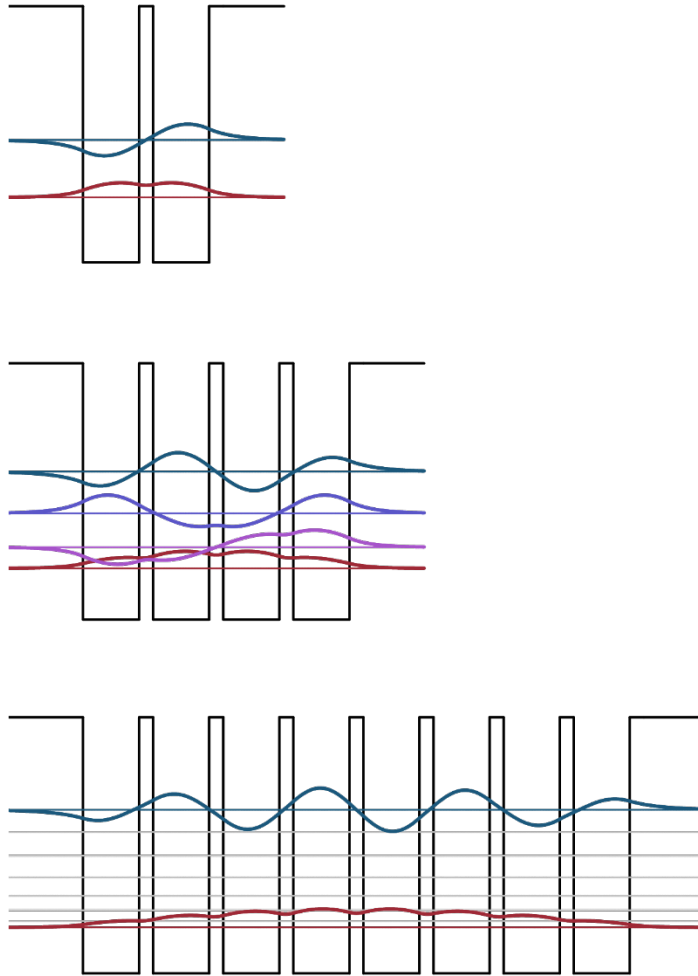


Fig. 4.11. Illustration of the coupling of multiple identical wells. The calculations are for an electron with a potential barrier height of 1 eV, well thickness of 0.6 nm, and barrier thickness between the wells of 0.15 nm, for 2 coupled wells (top), 4 coupled wells (middle), and 8 coupled wells (bottom). The wavefunctions of just the lowest and highest energy states are shown for the case of the 8 coupled wells.

³⁴ These calculations are also performed using the transfer matrix technique.

³⁵ We are using different barrier thicknesses between the wells in these different figures, but the well widths and barrier heights do correspond.

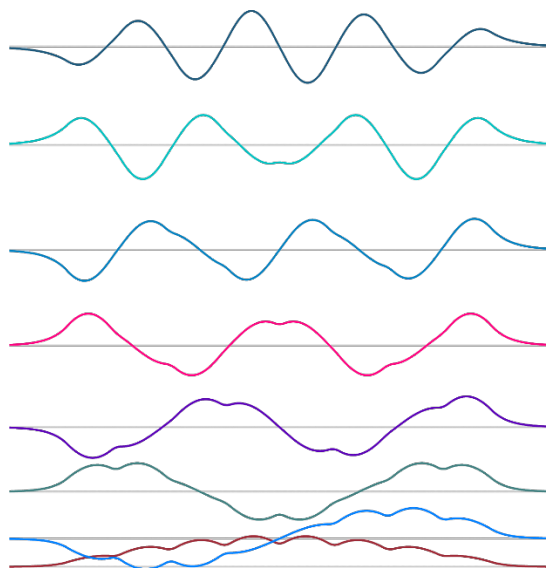


Fig. 4.12. Expanded detailed view of the wavefunctions for the case of eight coupled wells as in Fig. 4.11.

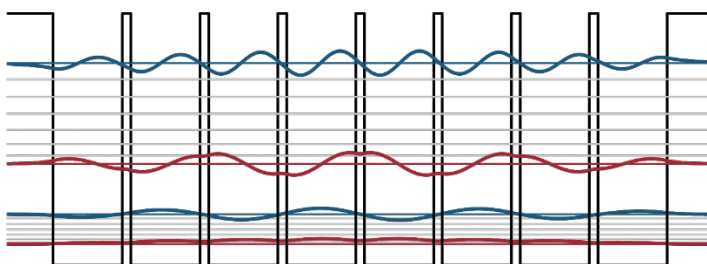


Fig. 4.13. Coupling of multiple wells, with well width 1.2 nm, barrier width 0.15 nm, and barrier height 1 eV, showing two bands of states, each corresponding to a different state in the isolated well. (See the bottom panel in Fig. 4.8 for the wavefunctions and energies of such an isolated well.) The wavefunctions for the lowest and highest states in each band are shown.

As we increase the number of coupled wells or “atoms”, we call the resulting groups of coupled states “bands”. Notice that, as we increase the number of coupled wells, the separation between the lowest energy and highest energy coupled states increases, but actually that separation starts to “saturate” with increasing numbers of wells³⁶. So in the bottom panel of Fig. 4.11 we are seeing a band of 8 states, all of different energies, with the separation between the lowest and highest energies in a band – the energy width of the band – tending towards a limit as we increase the number of wells. Further increases

³⁶ In one of the simpler approximate approaches, the “tight-binding” model, for example, the energy width of the band limits to twice the separation of the two states in a coupled well.

beyond these 8 wells here would lead to relatively little increase in this energy separation between the lowest and highest energy states in the band.

If we increase the width of the wells being coupled, the energy of the first level will decrease and we can usefully look also at what happens to the second level. This is shown in Fig. 4.13 where we have doubled the thickness of the individual wells³⁷. The details of the wavefunctions for the lower energy band in Fig. 4.13 are shown in Fig. 4.14, and similarly the wavefunctions for the upper energy band in Fig. 4.13 are shown in Fig. 4.15.

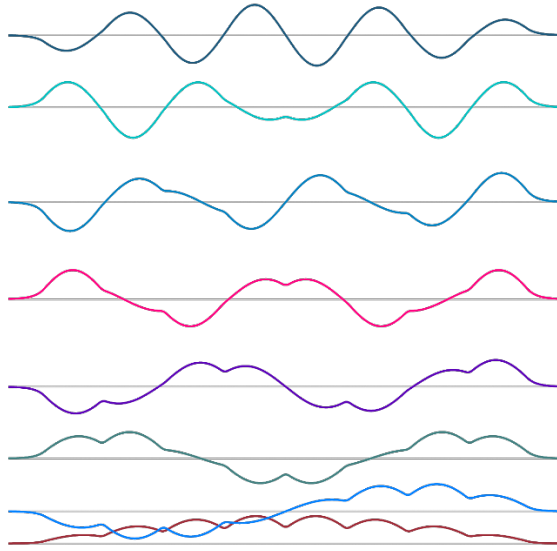


Fig. 4.14. Expanded detailed view of the wavefunctions for the lower energy band for eight coupled wells as in Fig. 4.13.

If we look at these sets of wavefunctions for the different levels within a band, we can start to see a pattern emerging. These wavefunctions are a product of (i) a “unit cell” function that is essentially the “isolated well” function or is closely related to it, and (ii) an overall standing-wave “envelope” function. For example, for the lowest state in the band in Fig. 4.11 and Fig. 4.12 for the case of 8 wells, we can see an overall standing-wave “envelope” pattern corresponding to one half-wave of a sine wave, just like a wave on a string or a particle in a box, that is “modulated” by one “bump” per well.

For the highest state in this band, we see that the wavefunction is changing sign between adjacent wells. This is because the “unit cell” or isolated-well function is multiplied by an envelope that has 4 full waves (or 8 half-waves) between the “walls” at either end. For the other levels between these two, if we look closely we can discern that the “envelope” function for each successive higher energy state has one more half-wave in it. We see the same behavior for the lower energy band for the eight coupled wells in the wider well case shown in Fig. 4.13 and Fig. 4.14.

³⁷ The parameters for this wider individual well here are the same as in the bottom panel of Fig. 4.8.

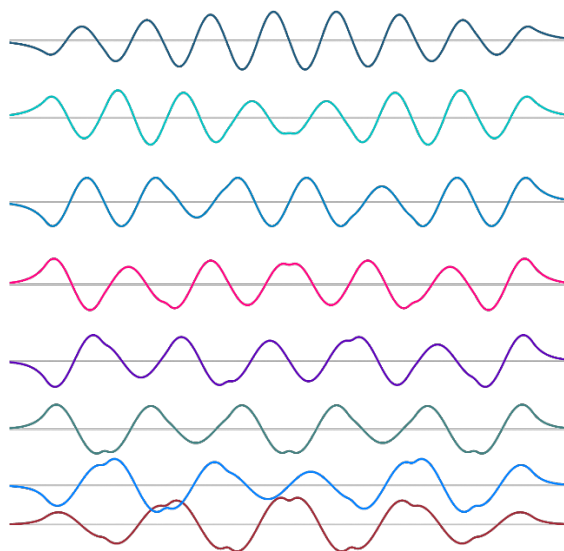


Fig. 4.15. Expanded detailed view of the wavefunctions for the upper energy band for eight coupled wells as in Fig. 4.13.

The upper band in Fig. 4.13 and Fig. 4.15 is showing similar behavior, with a “unit cell” or “isolated well” wave function multiplied by a similar envelope. We notice two differences of detail here compared to the lower band: (i) the “unit cell” or “isolated well” wavefunction itself passes through zero somewhere in the middle of each well, just like the second state in the isolated well; (ii) the order of the envelope functions is reversed, with the “single half-wave” one being at the top of the band, and the “8 half-waves” one being at the bottom of the band – this different order is, however, consistent with the amount of “curvature” in the overall waves being larger in the higher energy states.

We can understand why the energy width of the band saturates. We can see in that, at one extreme of the band energies, the envelope function will tend towards a very flat form as we increase the number of “periods” in our finite structures; then, nearly all the curvature, and hence the “kinetic energy” associated with the wavefunction, comes from the “unit cell” part, so this “kinetic energy”, and hence the overall energy of the state, will saturate out to some value as we increase the number of periods. Similarly, at the other extreme of the band energies, the envelope function tends towards a sinusoid that changes sign between each adjacent period, and again the overall “curvature” and hence the “kinetic energy” will saturate out to some value. Hence, as we increase the number of periods in these structures, we expect the band energy width to saturate out to some value that becomes independent of the number of periods.

Though shown here for a finite number of periods (or “atoms” or wells), the kinds of behaviors discussed in this section continue as we increase the number of periods and hence as we consider crystals more generally. If we allow ourselves one mathematical simplification, we can express this behavior particularly elegantly through what is called the Bloch theorem.

4.13 Band structures in crystals

A crystal will generally contain a very large number of atoms. A small piece of material you might pick up between your fingers could have many more than 10^{20} atoms in it, for example. This is obviously much larger than the 8 or so periods we have been considering in our direct calculations so far. Directly calculating the properties of such a large collection of atoms would be quite impossible.

Crystals, however, are naturally very regular, and we might hope that regularity would simplify the problem. Two particular approaches – the one-electron approximation and the Bloch theorem – when used together do indeed give us this simplification and some very useful models. These lead to “band structures” and to the widely used idea of effective mass for electrons. We introduce all these ideas here.

One-electron approximation

When we were discussing many-electron atoms above, we already introduced a key idea that we will extend here. This is the idea that, in a complicated system with very many electrons and nuclei, we may be able to pretend, at least as a useful approximation, that an electron sees an average potential from all the other electrons and nuclei³⁸.

In the case of a crystal, which is a structure that is periodic, we can also pretend that average potential $V_p(\mathbf{r})$ is itself also periodic, with the same periodicities and symmetries as the crystal itself. For example, for a one-dimensional system with a repeat length of a , we would have

$$V_p(z+a) = V_p(z) \quad (4.31)$$

We can keep on extending this for subsequent periods; that is,

$$V_p(z+2a) = V_p(z+a) = V_p(a) \quad (4.32)$$

so quite generally we can write for our “periodic potential”

$$V_p(z+ma) = V_p(z) \quad (4.33)$$

where m is an integer. We could also appropriately generalize this for two-dimensional or three-dimensional periodic structures and corresponding periodic potentials. Hence we presume that, at least as a first approximation, for any given electron, we can write a Schrödinger wave equation for just that one electron in the form

$$-\frac{\hbar^2}{2m_o} \nabla^2 \psi(\mathbf{r}) + V_p(\mathbf{r}) \psi(\mathbf{r}) = E \psi(\mathbf{r}) \quad (4.34)$$

³⁸ Again, an approach like this is sometimes called a Hartree-Fock approximation.

where $V_P(\mathbf{r})$ is periodic in three dimensions³⁹, just like the crystal. This is what we call the one-electron approximation for our crystalline structure.

This is certainly an approximation. There are many phenomena that are not covered by such an approach, such as the scattering of electrons off one another or from the nuclei (or vibrations of them), both of which are significant causes of electrical resistance in materials, for example. Often, though, we can model such phenomena, if they are not too strong, using this one-electron model as a starting point and adding in these other effects as what we call “perturbations”.

One of the most important consequences of this model is that it allows a simple way of looking at crystals, through the Bloch theorem.

The Bloch theorem 14b

In the calculations above for structures with specific numbers of periods, we observed some emerging behaviors; specifically, the eigenfunction solutions start to look like a “unit cell” function, which is the same in every period, multiplied by a “standing wave” envelope. For these finite structures, the exact positions for where the zeros would be near the ends of the standing wave are not quite simply defined – they will depend on the energy of the states to some degree. We could make some approximation there, and thereby construct a viable model for systems with large numbers of periods based on such a “standing wave” approach, and that can certainly be done.

A much more common approach, however, is to pretend we are solving a different but related problem, one that has exact and elegant mathematical properties. Specifically, instead of considering a “chain” of “unit cells” with fixed ends, we pretend that we join the two ends of the “chain” together into a loop or ring, as illustrated in Fig. 4.16.

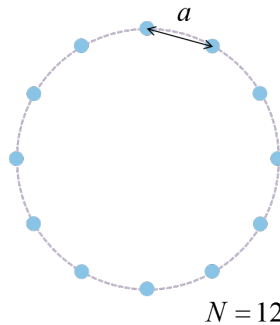


Fig. 4.16. Illustration of a chain of 12 “atoms” or unit cells joined in a ring.

With z as the distance along this loop, then on this loop, the potential can be written

$$V_P(z + ma) = V_P(z) \quad (4.35)$$

³⁹ If we were dealing with a two-dimensional “sheet”-like structure, then we could have a similar equation where $\mathbf{r}$ was a vector in two dimensions.

where m is an integer⁴⁰. This expression for the “periodic potential” V_p in Eq. (4.35) is just like the one for the infinite crystal in Eq. (4.33). If this chain is very long, we can argue that its internal properties will not be substantially different from those of an infinitely long chain, so this is a good model that gives us a finite system while keeping it periodic.

There are some quite real problems for which such an approach is rigorously correct; an example would be a “benzene ring”, which is exactly such a ring of six carbon atoms⁴¹. In the physics of larger crystals, though, this presumption that we can pretend the ends of the crystal are joined round onto one another is just to make the mathematics simpler.

The rationalization as to why we can get away with this for crystals is that, if the chain is sufficiently long, we are arguing that its properties cannot really depend much on whether the ends of the chain are joined together. Of course, simultaneously joining each opposite pair of surfaces of a three-dimensional object, like a “cube” of some physical crystalline material, to one another is geometrically absurd. Nonetheless, we do this mathematically anyway for our models because it makes the mathematics so much simpler. This approach is called using “periodic boundary conditions”. It does work in practice very well for crystalline structures that are at least moderately large – for example, perhaps at least 100’s of atoms in each direction⁴², and we may be able to use it for somewhat smaller structures also.

In periodic boundary conditions, then, we imagine we have a chain of N unit cells that are joined in a ring. If we presume the length of the unit cells is a , then the length of the entire chain is Na . We do want the wavefunction to be single-valued; otherwise how could we differentiate it or evaluate its squared modulus, for example? The periodic boundary conditions then require that, if we go all the way round the ring, the wavefunction solutions must get back to where they started. So for some wavefunction solution $\psi(z)$, where z is the distance along the chain, this requirement that the wavefunction gets back to where it started when we go all the way round the ring (the “periodic boundary conditions”) implies that, for any z ,

$$\psi(z + Na) = \psi(z) \quad (4.36)$$

For the case of quantum mechanical waves, we also want any measurable quantity, such as the electron density, to be the same in every unit cell, so we want

$$|\psi(z + a)|^2 = |\psi(z)|^2 \quad (4.37)$$

which means that the electron density in one unit cell has the same form as the one in the next unit cell, and so on all the way round the chain. We can satisfy both of these conditions, Eqs. (4.36) and (4.37), if we propose wavefunction solutions of the form

⁴⁰ m can even be larger than N and this will still work.

⁴¹ Each carbon atom also has one hydrogen atom attached to it in a benzene ring.

⁴² If we have to analyze a small structure, then we will have to be careful, and we may have to stop using periodic boundary conditions, reverting to actual “hard wall” boundary conditions at the ends of some finite structure.

$$\psi(z) = u(z) \exp(ikz) \quad (4.38)$$

where $u(z)$, the “unit cell function”, is the same in every unit cell – that is

$$u(z + a) = u(z) \quad (4.39)$$

and where k takes on any of a set of N values that are spaced by an amount $2\pi / Na$, a set that is conventionally written in a “symmetric” form⁴³ as

$$k = \frac{2n\pi}{Na}, \quad n = 0, \pm 1, \pm 2, \dots, \pm \frac{N}{2} \quad (4.40)$$

The assertion that we can write the solutions for such periodic boundary conditions in the form of Eq. (4.38) with such restrictions on k is known as the Bloch theorem. It is straightforward to check that this form does indeed satisfy both conditions Eqs. (4.36) and (4.37). In this form, the “ $\exp(ikz)$ ” function is an example of an “envelope” function that multiplies the unit cell function, $u(z)$.

The main mathematical differences between this Bloch form solution, Eq. (4.38), and the “standing wave” solutions for a finite length of crystal are that (i) the Bloch form corresponds to traveling, not standing, waves, (ii) the finite crystal with “ends” to it has standing wave “envelopes” that correspond to fitting integer numbers of half-waves between the ends, whereas these Bloch form solutions fit integer numbers of whole waves round the ring, but the propagating waves in opposite directions count as different solutions, so we get to the same number of different solutions overall^{44,45}, and (iii) such solutions in this Bloch form will in general be complex.

We can formally return to our sets of potential wells and solve for the states of them now using the periodic boundary conditions⁴⁶. We can choose the “left” of our unit cell at some position z_L (for example, in the middle of the barrier on the left) and the right of the unit cell therefore at some position $z_R = z_L + a$, where a is the repeat length (so

⁴³ Note that writing these allowed values this way (i) presumes N is even (though that is not a necessary restriction for using periodic boundary conditions), and (ii) leads to $N+1$ elements in the list. These “details” are customarily ignored because N is presumed large enough that no one cares much about the difference between N and $N+1$ or whether it is even or odd. To get the right number of elements, we can leave off either $-N/2$ or $+N/2$. An alternative list that avoids both of these minor issues is to choose a “single-sided” form $n = 0, 1, 2, \dots, N-1$, though, by convention, we typically do not use this list, mostly because it is not symmetric about zero. Formally, we are mathematically writing out the N different “ N th roots of unity” – the N distinct numbers that, when raised to the power N give the result 1.

⁴⁴ Except, of course, that, with k values as stated in Eq. (4.40), we have one too many in the Bloch form case.

⁴⁵ One more substantial “error” in the Bloch form is that it allows a $k = 0$ solution that is not quite physical for a crystal of actual finite length; that solution would be zero everywhere in the “standing wave” case if the wavefunction has to hit zero at the walls.

⁴⁶ By using the Bloch form of the solutions, this case of “rectangular” potential wells solved with such periodic boundary conditions can be completed analytically (though the solutions are not in closed form). That particular analytic approach is known as the Kronig-Penney model.

z_R would then be in the middle of the barrier on the right). Now, from Eq. (4.38) with Eq. (4.39), we have⁴⁷

$$\begin{aligned}\psi(z+a) &= u(z+a)\exp[ik(z+a)] = u(z)\exp(ikz)\exp(ika) \\ &= \exp(ika)\psi(z)\end{aligned}\quad (4.41)$$

Solving with periodic boundary conditions

Then, in this Bloch form approach, presuming we know the periodic potential, $V_P(\mathbf{r})$ (which becomes just our “rectangular well” potential in the z direction in our example here), our mathematical task formally becomes solving the Schrödinger equation, Eq. (4.34), subject to the boundary conditions

$$\psi_{q,k}(z_R) = \exp(ika)\psi_{q,k}(z_L) \quad (4.42)$$

Now we are being explicit in our notation for the wavefunction $\psi_{q,k}(z)$ that we are solving for a wavefunction in a specific “band” q and at a specific one of the allowed values of k .

For each k we expect to get a set of eigen solutions, indexed by some integer q that we could call the “band index”, with eigenfunctions

$$\psi_{q,k}(z) = u_{q,k}(z)\exp(ikz) \quad (4.43)$$

and some corresponding set of eigenenergies $E_{q,k}$. Effectively, this approach means that for each allowed k we are really solving for the unit cell function, $u_{q,k}(z)$, and the associated eigenenergies. The results of one such calculation for our periodic set of “rectangular” potential wells are shown in Fig. 4.17.

We see from the top trace in Fig. 4.17 that the unit cell function for this particular k value (or for this “ k state” in the band) is indeed periodic with the periodicity of the “crystal” – that is, it is the same in each unit cell⁴⁸. The unit cell function for the state in this lower band has no zeros inside each well, as we might expect since it is formed from the lower state in the isolated potential well.

The middle trace shows the (real part of) the propagating wave “envelope” function. For this particular chosen value of k , the “envelope” propagating wave goes through one full cycle from the left to the right of the figure. The bottom trace shows the (real part of) the total wavefunction, which is the product of the unit cell function and the propagating wave “envelope” function, as we expect for such a Bloch form of solution.

⁴⁷ This form here, $\psi(z+a) = \exp(ika)\psi(z)$, is just as good a statement of the Bloch theorem result as are Eqs. (4.38) and (4.39), and is often used mathematically instead.

⁴⁸ Or, at least the real part, as shown, though in fact the complete complex unit cell function is also the same in each unit cell.

The total wave in Fig. 4.17 corresponds here approximately with the second state from the bottom in Fig. 4.14⁴⁹.

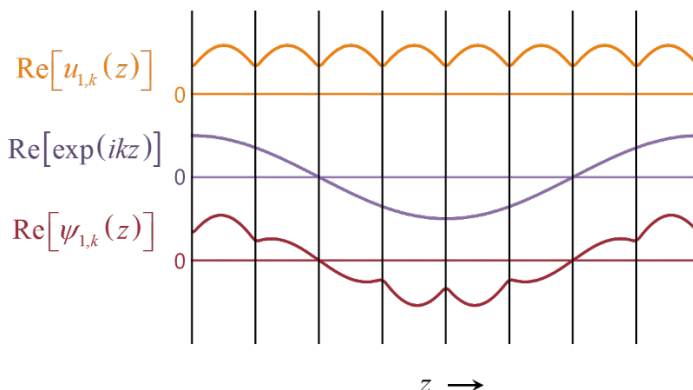


Fig. 4.17. Calculation of the wavefunctions for a periodic set of potential wells, each like those in Fig. 4.13 (well width 1.2 nm, barrier width 0.15 nm, giving a repeat length $a = 1.2 + 0.15 = 1.35$ nm, and barrier height 1 eV), and for a k value of $0.25\pi / a \approx 0.582 \text{ nm}^{-1}$. The vertical lines are drawn at the positions of the middles of the barriers between adjacent wells. The real parts of the complex wavefunctions are shown. The calculation is for one state (the second lowest) in the lower energy band, which corresponds to coupling the “lower” states in each well.

Essentially, what the Bloch theorem approach has done is transformed a relatively intractable problem of solving for all the states of a large system of many atoms into a problem that is roughly like solving for the states of just one “unit cell”, which is more like solving for the states of just one atom. This problem is somewhat more complicated than the hydrogen atom problem because (i) we have to guess what the potential energy is, possibly iterating to get some form so the calculated energies agree with measured energy separations in the material, and (ii) we have to solve the problem again for each value of k (though just with the same potential). Such “band structure” calculations are a major branch of the subject of solid-state physics.

Quiz questions

Q4.13.1 For the effective potential $V(z)$ seen by an electron in a one-dimensional crystal with repeat length a (in the z direction) state for each of the following whether it is necessarily true:

- a) $V(z) = V(z + 2\pi a)$
- b) $V(z) = V(-z)$

⁴⁹ The propagating wave envelope function in Fig. 4.17 has a different phase compared to the standing wave in Fig. 4.14, so the total wave is shifted laterally in Fig. 4.17, though that phase is arbitrary, so there is no particular meaning to this lateral shift.

c) $V(z) = V(z + 27a)$

Q4.13.2 We connect 15 atoms, regularly spaced by a distance a , in a circular “chain” or ring. The variable z can be taken as the distance around the chain in a clockwise direction, relative to the center of one of the atoms, which we put at top-center of the ring. Here, $V(z)$ is the effective potential energy seen by an electron, $\psi(z)$ is the electron wavefunction corresponding to some energy eigenstate in this potential, and k is presumed to be one of the values of k allowed for the Bloch theorem. For each of the following statements, state whether is necessarily true (including being true for any of the energy eigenstates in the case of wavefunctions).

a) $V(z) = V(z + 3a)$

b) $\psi(z) = \psi(z + 5a)$

c) $|\psi(z)|^2 = |\psi(z + 7a)|^2$

d) $\frac{2\pi}{3a}$ is an allowed value for k .

e) $\frac{\pi}{2a}$ is an allowed value for k .

f) 0 is an allowed value for k .

Q4.13.3 For a wavefunction $\psi(z)$ that is in Bloch form for a one-dimensional crystal with repeat length a , where k in each case is any of the corresponding allowed k -values in such a crystal, and $u(z)$ is some appropriate function for which $u(z) = u(z + a)$, state whether each of the following is necessarily true.

a) $\psi(z) = u(z) \exp(ikz)$

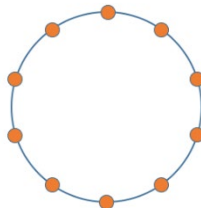
b) $\psi(z + a) = \exp(ika) \psi(z)$

c) $\psi\left(\frac{z}{2} + a\right) = \exp\left(i\frac{k}{2}a\right) \psi\left(\frac{z}{2}\right)$

d) $\psi(z + a) = u(z + a) \exp(ik(z + a))$

Problems

4.13.1 Consider a ring of 10 atoms spaced by a distance $a = 5$ nm round the ring



- a) Write out all 10th roots of unity, r_n , such that $r_n^{10} = 1$. (Choose the way you write these so that the case of $n=1$ corresponds to the very simple root “1”, and try to proceed sequentially with n as you write down the subsequent roots. You should actually be able to write a simple formula for these roots as a function of n from 1 to 10, and you can leave the result in some exponential form, for example.)
- b) Considering the wavefunction of an electron as a function of the distance s round the ring, write down the periodic boundary condition for the wavefunction as a function of s for this 10-element ring (using units of nm for the distance). Write down the corresponding relation between the modulus squared of the wavefunction at a point s_o and a point $s_o + 5$ nm.
- c) By considering the Bloch form of the solutions of such a problem (that is, a product of a unit cell function and a propagating wave “envelope” function with appropriate k values), calculate all the distinct possible values of k vectors that are allowed for the Bloch “envelope” functions. (Note: use a “single-sided” list of allowed values of k here, not the more common “symmetric” list of Eq. (4.40), so corresponding to running from 0 to $N-1$ rather than running from $-N/2$ to $+N/2$.)
- d) If the electron is in the state with the lowest allowed k vector magnitude, what is the probability of finding the electron in the first unit cell (the first unit of length a round the ring, starting from the top of the diagram and going to the right)? What is the probability of finding it in the sixth unit cell (the one just to the left of the bottom of the diagram)? How do your answers change if you consider the electron in the state with the fourth lowest k vector magnitude?

[Note: if you conclude that some aspects of this question are actually trick questions, you might be right!]

4.13.2 For a wavefunction $\psi(z)$ that is a solution of Schrödinger’s time-independent equation for an electron wave in a periodic potential with period a in the z direction, we must have that $\psi(z+a) = \psi(z)$. True or False? (Justify your answer briefly.)

4.13.3* Consider a system with some specific periodic potential where $V_P(z+a) = V_P(z)$. Suppose we know that an eigenstate of such a system is

$$\psi(z) = \left[\exp\left(-i\left(\frac{2\pi}{a} - k\right)z\right) - \exp\left(i\left(\frac{2\pi}{a} + k\right)z\right) \right]$$

Rewrite the eigenstate in the Bloch form $\psi(z) = u(z)\exp(ikz)$, where $u(z)$ is a function that is the same in every unit cell. [Hint: think about how trigonometric functions can be written using complex exponential functions.]

4.13.4 Suppose that, in a periodic potential of period a in one-dimension, a function $\psi_1(z) = u_1(z)\exp(ik_1z)$ where $u_1(z) = u_1(z+a)$ is known to be a solution of the

one-electron time-independent Schrödinger equation under periodic boundary conditions, with some specific eigen energy E_1 . Consider now the function $\psi_2(z) = u_2(z) \exp(ik_2 z)$ where $k_2 = k_1 + 2\pi/a$. Give an expression for $u_2(z)$ that will guarantee that $\psi_2(z)$ is also a solution of the same Schrödinger equation with the same eigen energy E_1 .

4.14 Band structures in one dimension 15a

We can continue our example calculations for this periodic structure of “rectangular” potential wells. Specifically, we can plot all the eigen energy solutions for each one of the possible values of k as given by Eq. (4.40). For $N=8$ unit cells in the periodic structure, we get the set of energies given by the dark dots in Fig. 4.18.

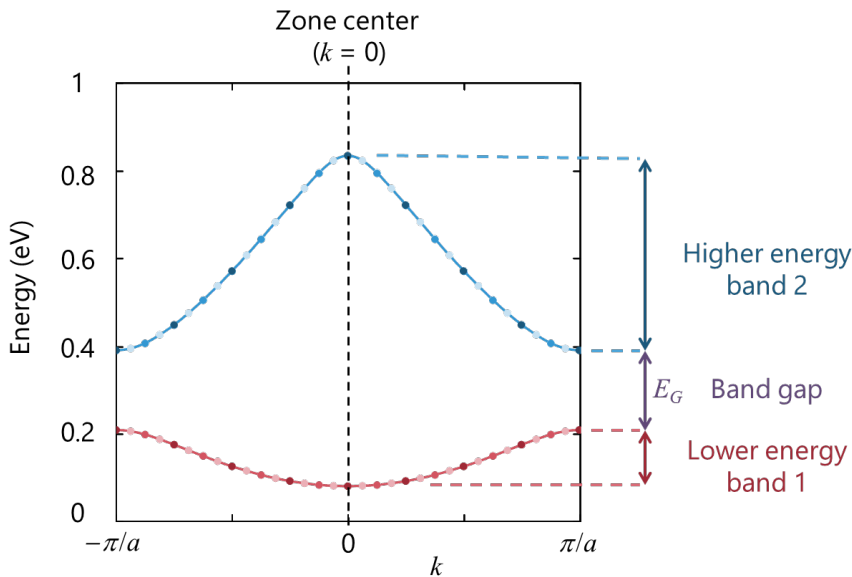


Fig. 4.18. Band structure (eigenenergies as a function of wavevector k) for a periodic structure of rectangular wells (well width 1.2 nm, barrier width 0.15 nm, so a repeat length $a = 1.2 + 0.15 = 1.35$ nm, and barrier height 1 eV). The dark-colored dots are for $N=8$, the medium-colored dots show the additional results for $N=16$, and the light dots show the additional results for $N=32$. The solid lines show the limiting behavior for very large numbers of periods.

First, note that the k values for these dots are equally spaced, by an amount $2\pi/Na$, and, consistent with Eq. (4.40), these run from $-\pi/a$ to $+\pi/a$. To get the correct number (8) of distinct states, we could just leave out either one of the “end points” – that is, either the $-\pi/a$ or the $+\pi/a$ value for k . We can see that, as we have “pushed” the “atoms” together to make this “crystal”, we can keep the same counting of states. That is, some given atomic level in each of N atoms has turned into N energy levels in

a band in the combined system, with the same such counting behavior for each atomic level.

Second, note that, for each value of k , within the energy range from 0 to 1 eV for which we performed calculations here, there are two possible energy eigenvalues, as given by the “red” dots for the lower band and the “blue” dots for the upper band.

Third, as we increase N , here by doubling N first to 16, and then to 32, we end up merely “filling in” the spaces between the existing “dots”, retaining all the previous solutions. The solid lines show the limiting cases for large N . Note in particular that the energy width of the bands does not increase as we keep on increasing N to arbitrarily large numbers. We end up with energy bands of definite width, as shown for the lower and higher energy bands 1 and 2 in Fig. 4.18. Note also that these bands are smooth and continuous once we increase to large N , and all the allowed energy solutions lie on these curves.

Fourth, we see in this case that there is a range of energies between the two bands for which there are no energy eigenstates. Such energy ranges are called “band gaps”, with a band gap energy E_G . The existence or otherwise of band gaps in a band structure can be very important in influencing both the electrical and optical properties of a material.

Here we have only calculated for the allowed set of N different k values. Mathematically, we could keep on performing calculations for further values of k spaced by further amounts $2\pi / Na$. Essentially, we would just be continuing the band structure of Fig. 4.18 horizontally. If we did that, what we would find is that we would just keep repeating the same energy solutions; we could just take the curves in Fig. 4.18 and copy them sideways, shifting them over by an amount $2\pi / a$ each time. The total wavefunctions for these solutions would also just be the same as for the corresponding solution in the diagram in Fig. 4.18. Any one such $2\pi / a$ range of k values can be called a “Brillouin zone”. Conventionally, the diagram centered round $k = 0$, like the one in Fig. 4.18, is referred to as the first Brillouin zone, and the point $k = 0$ is referred to as “zone center”. We only ever need to calculate the results for the first Brillouin zone because it contains all the distinct solutions.

Crystal momentum 15b

For a particle in a uniform potential, we know that the eigen solutions take the form of plane waves. For example, in one dimension, for a particle of mass m_b with a given energy E , solving Schrödinger’s equation would give a wavefunction $\propto \exp(ikz)$ for a “forward-going” wave, with corresponding eigen energy $E = \hbar^2 k^2 / 2m_b$. We also know from de Broglie’s hypothesis that the magnitude of the momentum for such a situation is $p = \hbar k$.

Now, in the case of these solutions in Bloch form, we also have as the “envelope” part of the solution a propagating wave of the form $\exp(ikz)$. We can if we want define a similar concept here that we call the “crystal momentum” p_C , which we write down with a formula that looks essentially the same as de Broglie’s hypothesis,

$$p_C = \hbar k \quad (4.44)$$

where the k here is the k in the propagating “envelope” function.

This quantity turns out to have many effective properties that behave like actual momentum for an electron in a crystal. In many ways, it can appear as if an electron in the crystal has such a momentum. We should be clear, though, that, because we are only considering the “envelope” wavefunction of the form $\exp(ikz)$ here rather than the full wavefunction of the electron, p_C is not actually the momentum of an electron in the crystal; we could call this effective momentum a “pseudo-momentum”.

It can, for example, appear that an electron accelerated by a force in the crystal acquires just such an effective momentum, and in processes like optical absorption, it can appear that “momentum” overall is conserved when we use this pseudo-momentum as the effective momentum of the electron. In all these cases, in fact we deduce this effective conservation of this pseudo-momentum from deeper analysis of the actual problem of interest – the effective “conservation” of this pseudo-momentum is something that comes out of that analysis, not a principle we put into the analysis. Nonetheless, this concept turns out to be so useful in sufficiently many situations that it is common in discussions of electrons in crystals to refer, loosely, to this crystal momentum as the electron momentum. One particular example of this occurs when we consider effective mass.

Effective mass

We have seen already in our example calculations for the case of a periodic structure of “rectangular” wells that we get a band structure that has minima and maxima at particular points in the Brillouin zone (that is, for particular values of k). Near some minimum or maximum, at least for some small range of k , we could fit the energy as a function of k with a parabola, as sketched in Fig. 4.19.

For such a parabola centered round $k=0$, we could write down an approximate formula for the electron energy, relative to the energy at the minimum or the maximum, E_o , that, by definition of a parabola, would have the form $E - E_o \propto k^2$. One convenient way of writing down the proportionality constant is to write this relation in the form

$$E - E_o = \frac{\hbar^2 k^2}{2m_{eff}} \quad (4.45)$$

This parameter m_{eff} , which we call the “effective mass”, is then just a parameter with dimensions of mass that characterizes how shallow or steep this parabolic variation of energy is near to $k=0$. A large value of m_{eff} – a large “effective mass” – corresponds to a shallow parabola in Fig. 4.19, and a small value – a small “effective mass” – corresponds to a steep parabola.

If we then treat $p_C = \hbar k$ as an effective momentum, then we get a relation between the energy $E - E_o$ and this effective momentum that is of the form

$$E - E_o = \frac{p_C^2}{2m_{eff}} \quad (4.46)$$

which is consistent with this effective momentum behaving in many ways like momentum for a particle of an effective mass m_{eff} . (For minima or maxima centered round some different value k_o of k , as in the conduction band case on the right of Fig. 4.19, we would replace k with $k - k_o$ in Eq. (4.45).)

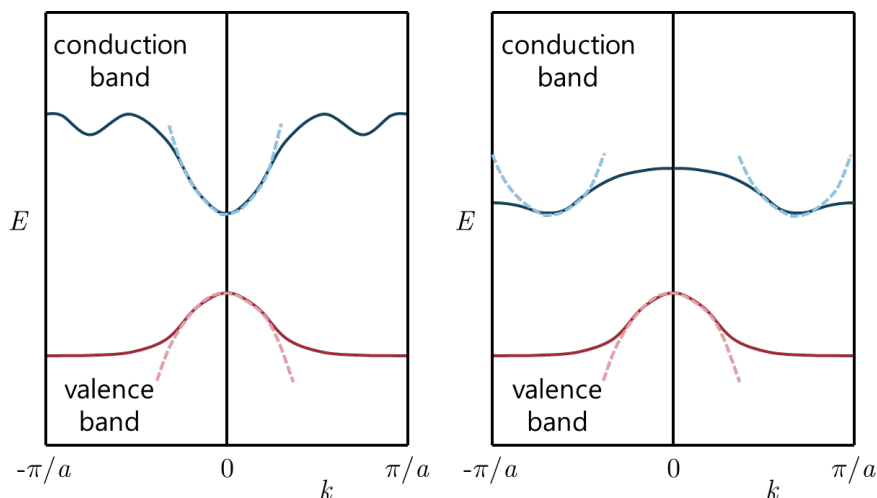


Fig. 4.19. Sketches of hypothetical band structures with a maximum in a lower “valence” band and a minimum or minima in an upper conduction band, together with parabolas (dashed lines) fitted to the maxima and the minima. The sketch on the left shows a “direct band gap” material, with the conduction band minimum and the valence band maximum at the same k . The sketch on the right shows an “indirect band gap” material, with the conduction band minima and the valence band maximum at different k .

This idea of effective mass is very useful, especially when discussing the behaviors of electronic and optoelectronic semiconductor devices. It is important to remember that it is only an effective concept, albeit a very useful one. We are not in fact changing the mass of the electron. But the band structures we get can usefully make it appear as if the electron mass is different.

Quiz questions

- Q4.14.1 For a larger crystal of the same material, because the atoms are still spaced by the same distance, the allowed k values for the Bloch functions are also still spaced by the same amount. True or False?
- Q4.14.2 The number of k values corresponding to a band is equal to the number of unit cells in the crystal. True or False?
- Q4.14.3 For each k value, there is only one possible energy eigenstate for an electron. True or False?
- Q4.14.4 For a one-dimensional crystal with a repeat length of a , if there is an energy eigensolution with some energy E_k associated with a particular value of k , then

there is one with the same energy associated with the value $k + 2\pi / a$. True or False?

Q4.14.5 The effective mass of an electron in a given band is necessarily the same throughout the Brillouin zone. True or False?

Q4.14.6 Depending on the material, the band and the position in the band, it is possible for an electron effective mass to be negative. True or False?

Q4.14.7 The crystal momentum at a given value of k in a one-dimensional crystal is the same in all bands. True or False?

Problems

4.14.1 For a direct band gap semiconductor, if the lowest energy in the conduction band lies at $k = 0$, there must also be a minimum in the electron energy in the highest valence band at $k = 0$. True or False? (Justify your answer briefly.)

4.14.2 In a band structure, one of the bands has a minimum at zone center. At a point in the Brillouin zone near that minimum at which the k vector has a magnitude of 0.5 nm^{-1} , the energy of an electron in this band is $+100 \text{ meV}$ relative to the energy in that band at zone center. Presuming we can approximate this band minimum by a parabolic variation of energy with k , what value of effective mass does this minimum correspond to, expressed in units of the free electron mass m_0 (that is, if you think the effective mass is $b \times m_0$, your answer would be b)?

4.15 Band structures in three dimensions

So far we have discussed the Bloch theorem explicitly only for a “one-dimensional” crystal that we can think of as a chain of unit cells. For two-dimensional or three-dimensional crystals, we simply presume that, at least for mathematical convenience, we can simply extend the ideas by writing

$$\psi(\mathbf{r}) = u(\mathbf{r}) \exp(i\mathbf{k} \cdot \mathbf{r}) \quad (4.47)$$

where the unit cell function is the same in every unit cell of the crystal (in all two or three directions), and the components of the wavevector $\mathbf{k}$ all obey relations similar to Eq. (4.40) with the appropriate “repeat lengths” a for each direction (which will be different for each direction if the unit cell has different lengths in each direction).

We can construct band structures for two- and three-dimensional crystals in similar ways to those discussed here for one dimension. Of course, such crystals can be more complicated than those in one dimension, with potentially much more complicated forms of unit cells; to plot out the band structure, we would need a four-dimensional diagram, with three dimensions to correspond to the components of a vector $\mathbf{k}$ in the Bloch form, and a fourth dimension to indicate the corresponding energies. Of course, we cannot readily do that graphically. As a compromise, band structures for such two- or three-dimensional crystals are usually just plotted along very specific directions, allowing a band structure diagram that is not very different at first sight from those we

have plotted for one dimension. The horizontal axis will now correspond to the magnitude of $\mathbf{k}$ along a specific direction. It is also common to plot the band structure for several different specific directions in one overall graph.

To understand the basic meaning of band structures for three-dimensional crystals, for example, we need first to understand the Brillouin zone in three dimensions. Now rather than being just a one-dimension range of one k , it needs to be a three-dimensional construction that shows the range of each of the three components of k . Different crystal structures have different forms of such three-dimensional Brillouin zones. The one corresponding to diamond or zinc blende lattices in the way it is usually drawn is shown in Fig. 4.20. Note that this Brillouin zone has a three-dimensional shape.

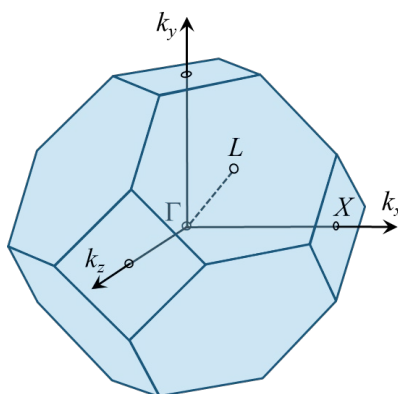


Fig. 4.20. The Brillouin zone for the face-centered cubic lattice, showing the ranges for the x , y , and z components – k_x , k_y , and k_z – of the $\mathbf{k}$ vector. The Γ (gamma) point is at $k = 0$, that is, in the center. The X direction is along a cube edge direction. The L direction is along the space diagonal direction of the cube.

Plotting band structures 15d

We might want to plot out the band structure along the x , y , and z directions, for example. Actually, for a crystal with such underlying cubic symmetry, we only need to do this along one of these three directions. The band structure along the x direction would look the same as that along the y direction or along the z direction.

Therefore, we could plot the band structures from the so-called Γ (gamma) point (the point at $k = 0$) out to the point X along the x direction, as shown in Fig. 4.20. Another interesting direction might be along the “space-diagonal” direction in the cube. In Fig. 4.20, that corresponds to the direction out to the L point. One other simplification is that band structures themselves are generally symmetric about the center of the Brillouin zone (the Γ point)⁵⁰. We see that symmetry already in the simple band structures we

⁵⁰ This symmetry is known as Kramers degeneracy. For each point on one side of the Γ point, there is one with the same energy symmetrically on the other side; technically, these are associated with opposite spin values, though states of opposite spin often have the same energy anyway, so band structures are often symmetric even for the same spin.

calculated for the one-dimensional lattice of rectangular wells, as in Fig. 4.18. As a result, we only need to plot in the “positive” half of the Brillouin zone.

When band structures are plotted, therefore, we can choose to use the “right” half of the plot for the band structure along one direction, such as the band structure out to the X edge point on the Brillouin zone, and the “left” half of the plot for the band structures along some other direction, such as out to the L edge point (that is, in the “space diagonal” direction).

Example forms of the band structures for silicon and for gallium arsenide are plotted in Fig. 4.21 and Fig. 4.22. In general, for band structures, we typically only plot some of the bands, usually including just the highest completely full bands (usually, valence bands) and the lowest completely or partially empty bands (usually, conduction bands).

Each band is formed from some combination of atomic states. The combinations associated with “core” electrons in the atom, while they do form bands, are usually of little technological importance, being still very tightly bound to the nucleus, and they may contribute relatively little to the bonding into a crystal form. So we typically do not plot those bands. We are most interested in the electrons in the outer, highest energy “valence” set of orbitals, just as we are generally in chemistry also.

For Group IV materials like carbon, silicon and germanium, in the outer, highest energy orbitals, we are dealing with the combination of 2 “s” orbitals and 6 “p” orbitals (where we include both spin states), which gives 8 orbitals that together hold 4 electrons. Putting atoms of one such type together to form a crystal leads to the formation of two sets of bands; at least loosely, one of these sets of bands is based on “bonding” versions of the combined orbitals of the different atoms and the other based on “antibonding” versions of those same orbitals. The bonding versions can have lower energy than the antibonding versions.

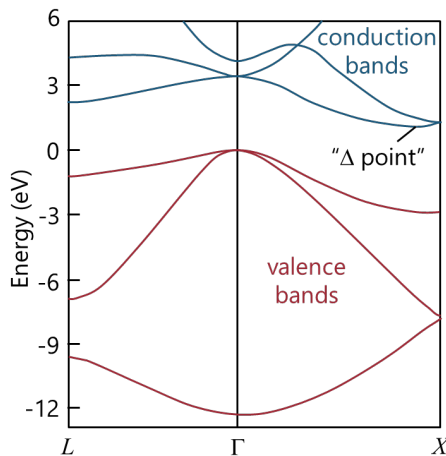


Fig. 4.21. Sketch of the upper valence and lower conduction bands for silicon, after calculations by K. S. Sieh and P. V. Smith, *Phys. Status Solidi* (b) **129**, 259 (1985).

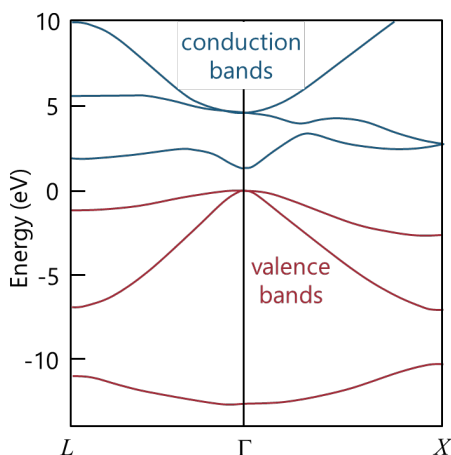


Fig. 4.22. Sketch of the upper valence and lower conduction bands for gallium arsenide, after calculations by M. Rohlffing, P. Krüger and J. Pollmann, Phys. Rev. B **48**, 17791 (1993).

If those “bonding” and “antibonding” sets of bands do not overlap in energy, the bonding versions will be essentially fully occupied, to give what are called the “valence” bands, and the antibonding versions will be essentially unoccupied, to give what are called the “conduction” bands⁵¹. Such non-overlapping band behavior, with a band gap energy between them, is characteristic of a semiconductor or an insulator⁵². In practical terms, the GaAs material and its various other III-V cousins behave essentially like effective Group IV materials, with similar behavior of the band structures.

We will return to discuss band structures and their consequences for physical properties and devices later. For the moment, we note a few aspects of these band structures for Si and GaAs. First we see that both of these band structures do show a band gap between the highest valence bands, which are all nominally fully occupied, and the lowest conduction bands, which are all nominally empty; this finite separation by a band gap energy is a necessary requirement for “semiconductor” behavior, which we will discuss in more detail later. By convention, the energy origin is taken at the top of the highest occupied valence band; this is just an arbitrary choice – the actual zero of energy there has no absolute significance.

In contrast to our simple one-dimensional coupled well cases, we have multiple bands here that do overlap in energy, and in some cases actually coincide with one another at specific points in the Brillouin zone. It is still the case, though, for all such band structures that, for N atoms in the crystal, each band has N states in it (multiplied by 2

⁵¹ For molecules, the terminology HOMO, for highest occupied molecular orbital, and LUMO, for lowest unoccupied molecular orbital, is often used.

⁵² We will return to discuss the difference between semiconductors and insulators, which is really a quantitative one based on the number of thermally excited electrons we have at normal temperatures; that number in turn depends on the band gap energy.

if electrons of different spins also have the same energy), and all bands have the same set of equally spaced k values associated with them.

The shapes of these bands are also more complicated than those of the simple one-dimensional “rectangular well” case. If the lowest point in the conduction band does not lie at the same point in the Brillouin zone (that is, at the same $\mathbf{k}$ value) as the highest point in the valence band, the material is said to have an “indirect” band gap (see also the right-hand diagram in Fig. 4.19), which is the case for silicon. By contrast, GaAs has a “direct” band gap, with the minimum in the conduction band lining up with the maximum in the valence band⁵³ (see also the left-hand diagram in Fig. 4.19). Whether a material has a direct or an indirect band gap greatly influences its optical properties, as we will see later.

This whole notion of band structures, which is mathematically enabled by the Bloch theorem (Eqs. (4.38), (4.39), and (4.40)), is very widely used in discussing a broad range of electrical and optical properties of crystalline solids, and therefore forms the scientific basis of much of modern electronic and optoelectronic device technology.

Quiz questions

- Q4.15.1 For an electron in a crystal, in an energy eigenstate associated with an allowed vector k value $\mathbf{k}$ in a crystal, the electron wavefunction can be written as $A\exp(i\mathbf{k}\cdot\mathbf{r})$ where A is a constant and $\mathbf{r}$ is the position in the crystal. True or False?
- Q4.15.2 The Brillouin zone for a face-centered cubic lattice necessarily has the shape of a cube when we plot it out as function of the three components of the vector $\mathbf{k}$. True or False?
- Q4.15.3 In a Brillouin zone in a three-dimensional crystal, there is necessarily one unique value of k that corresponds to the minimum possible energy for an electron in the conduction band. True or False?
- Q4.15.4 A minimum or maximum energy in a band is necessarily either at the center of the Brillouin zone or at an edge of the Brillouin zone. True or False?
- Q4.15.5 Some materials have their lowest energy minimum in the conduction band at the same value of $\mathbf{k}$ as the highest energy maximum in their valence band. True or False?

Problems

- 4.15.1 The band structures for three-dimensional crystals are usually just plotted along very specific directions because it requires a four-dimensional diagram to plot the energy E as a function of k_x , k_y , and k_z all on one diagram. Therefore, typically the band structures are plotted from the center Γ point out to the point

⁵³ In this particular case, the minimum and maximum are both at $k = 0$, though that is not a necessary requirement for a direct band gap

X along the x direction, and to the point L along the “space diagonal” direction, for the first Brillouin zone (Fig. 4.19 – Fig. 4.22) in a crystal with an underlying cubic kind of symmetry of the crystal structure.

- (a) By looking at the band diagram of silicon (Fig. 4.21), determine how many equivalent conduction band minima silicon has in its lowest conduction band. That is, how many distinct “valleys” at or near the Δ (delta) point(s) are there when we think about the crystal and the Brillouin zone in all three dimensions?
- (b) By looking at the band diagram of GaAs (Fig. 4.22), determine how many equivalent conduction band minima GaAs has in its lowest conduction band (considering only the lowest energy minimum or minima). In other words, at how many different points in the GaAs Brillouin zone in three dimensions will we find the same (lowest) minimum energy?
- (c) In the band diagram of GaAs (Fig. 4.22), we can see that the second lowest minimum in the conduction band as drawn here lies at the “ L ” point, at one edge of the Brillouin zone. How many equivalent such “second” conduction band minima are there for GaAs? [Hint: Think about what happens if a minimum “valley” is shared by two Brillouin zones.]

4.16 Electrons in bands

Now that we (or, in reality, someone else!) have calculated such a band structure, we can start to use it to understand some properties of crystalline materials. The basic approximation we now take is that, if we start to move electrons from one state to another, we can continue to use this same “one-electron” band structure to describe the states. That is, even though strictly we may have perturbed the perfect crystal by moving an electron from one k -state to another, including possibly putting it in a different band, or moving it to a different location in the crystal, we presume the underlying potential will remain approximately the same, including remaining approximately periodic. So, we can still use this band structure to model what happens when we move electrons around from state to state or place to place.

Again, this is certainly an approximation, but it turns out to work well enough for many purposes, and gives us a relatively simple way to think about crystalline materials and devices we might make from them.

Metals, insulators and semiconductors

One of the first things we can do with this kind of band structure approach is to understand the differences between various kinds of materials – in particular, metals, semiconductors, and insulators. Fig. 4.23 illustrates the basic difference in the band structures of insulators and semiconductors on the one hand, and metals on the other hand.

In insulators and semiconductors, the counting of the available valence electrons is such that, at least at low temperature, ideal perfect semiconductors and insulators have bands

that are either completely full or completely empty. The highest occupied bands are called the valence bands, and the lowest unoccupied bands are called the conduction bands. Both semiconductors and insulators have a finite band gap energy E_G separating the highest valence band and the lowest conduction band.

In metals, the counting of electrons and band states is such that there is a band that is partially filled with electrons, even at low temperature. Conventionally, that band is also called the conduction band⁵⁴. Metals may or may not have band gaps between various bands. Band gaps are relatively unimportant for metals because the electrical properties of metals are largely determined by the electrons in the partially filled conduction band. Band gaps are, however, crucial for the properties of semiconductors and insulators.

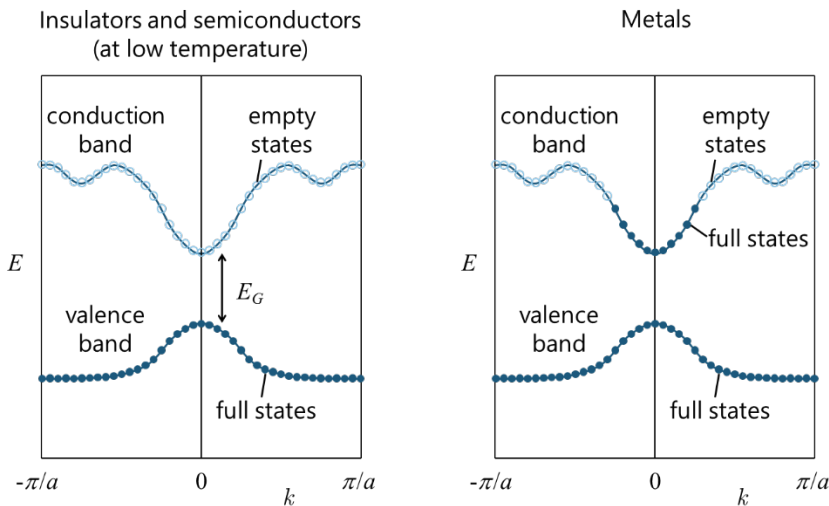


Fig. 4.23. Schematic illustration, in a one-dimensional case, of band structures (i) on the left, of insulators and semiconductors (at least at low temperature), and (ii) on the right, of metals.

In the case of a metal, with no applied electric field, of course we expect no current flowing. Indeed, for the kind of symmetric filling of the metal conduction band as shown in Fig. 4.23, there is no current. With a metal, if we apply an electric field, we will tend to “skew” the distribution of the electrons to one side or the other in our band structure diagram drawn as a function of k ; that “skewing” does correspond to a flowing current in one direction or the other⁵⁵. Of course, we can only skew the distribution if there are empty states into which to “skew” the electrons.

⁵⁴ Indeed, this idea of a conduction band only makes literal sense for metals. The “conduction” band does not necessarily “conduct” electrical current for semiconductors or insulators.

⁵⁵ We can think of this as unbalancing the momentum distribution of electrons, giving more with momentum in one direction than in the other. The actual calculation of currents from band structures is quite subtle, however, requiring the concept of group velocity to calculate it correctly. It is, though, still the case that this “skewing” of the electron distribution does correspond to current flow.

In the case of a semiconductor or an insulator, rather obviously if a given band, such as a “conduction” band, is empty, it cannot contribute to the current. Somewhat less obvious, however, is that a full band cannot conduct current either. Since all the states are full, we can make no changes in what states are occupied⁵⁶; we cannot “skew” the distribution of electrons, for example. So, with bands that are either completely full or completely empty, there is no current flow; hence an insulator can insulate.

Fig. 4.24 shows the bands plotted another way, still with increasing electron energy on the vertical axis, but with ordinary position on the horizontal axis. This kind of diagram does not show the k -states or the band structure as a function of k , but it is often used when discussing electronic and optoelectronic devices because those are typically made of multiple different materials joined together, and we need to see the behavior of band edges and carrier distributions especially near the junctions between the materials. On this diagram it is also clear that the insulator and the (low-temperature) semiconductor do not conduct electricity. Also in real space there is nowhere for the electrons to go in the valence band – all the states are already full – and obviously we cannot conduct in the “conduction” band because there are no electrons to give the conduction. For the metal, there are certainly available positions for electrons to move to, so conduction is possible.

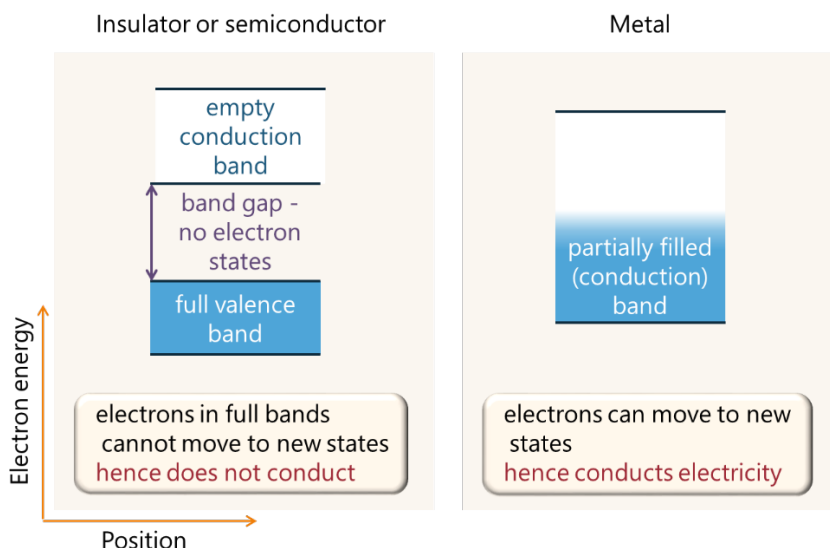


Fig. 4.24. Schematic illustration of bands in real space, in one direction (i) on the left, of insulators and semiconductors (at least at low temperature), and (ii) on the right, of metals.

To understand why a semiconductor can conduct electricity, and how we can achieve quite substantial control of that conduction to make transistors and other electronic

⁵⁶ Applying normal electric fields to semiconductors or insulators does not typically cause electrons to move from one band to another, though, with very high fields under some conditions, that kind of “interband” process is possible. This is the kind of mechanism that can then allow electrical breakdown of insulators.

devices, we need to understand first how finite temperature affects the precise distribution of electrons. Simply put, once we consider finite temperature, there can be small populations of electrons thermally excited across the band gap energy, and one result of that thermal excitation is that we can now conduct electricity, at least to some degree, in a semiconductor. The difference between semiconductors and insulators is essentially just the size of the band gap energy, which is relatively small for semiconductors, and relatively large for insulators. With a small band gap energy, that thermal excitation across the band gap is a stronger process, and hence such small band gap energies result in materials that partially conduct electricity (even when the materials are pure).

Electrons, holes, and doping

As materials with merely intermediate levels of conductivity, semiconductors would be relatively uninteresting electrically. What makes them much more interesting is that their electrical properties can be changed dramatically by the addition of small amounts of carefully chosen impurities called “dopants”. Here we will introduce some of the basic ideas and consequences of this introduction of impurities, known as “doping” (Fig. 4.25).

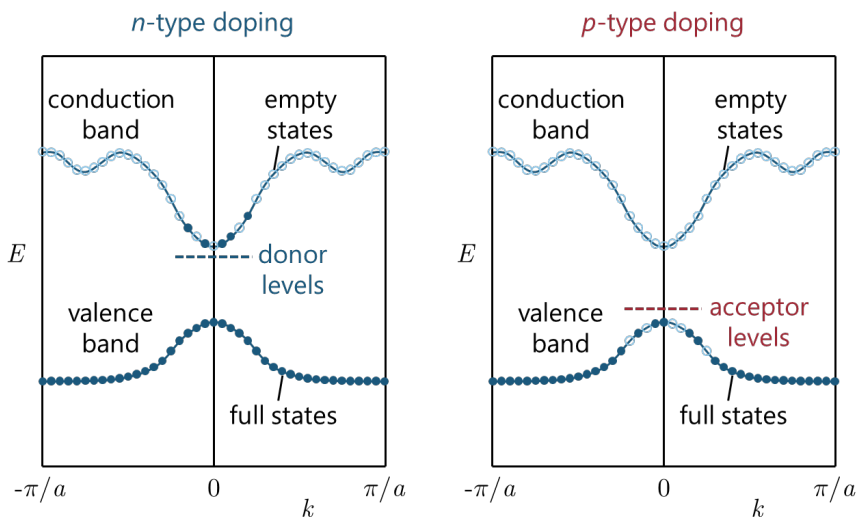


Fig. 4.25. Sketch of n -type doping (left) with added donor atoms giving donor levels just below the conduction band and leading to electrons ionized into the conduction band, and of p -type doping (right) with added acceptor atoms giving acceptor levels just above the valence band leading to “holes” ionized into the valence band.

This ability to control the amount and the nature of conduction in semiconductors by such controlled doping is central to the whole of semiconductor device technology, and is what enables all the major devices like diodes and transistors that make electronic technology possible. It is also central to the various optoelectronic devices, such as photodetectors, light-emitting diodes and lasers and many of the optical modulators that

turn light signals on and off in telecommunications. We will return to such points in later chapters.

The simplest view of doping of a Group IV semiconductor, like silicon or germanium, is that we substitute relatively small number of atoms, usually either from Group III or Group V of the periodic table, for some of the Group IV atoms in the crystal. Group V elements, such as arsenic or phosphorus, have an extra electron in their outer “shell” compared to the Group IV elements, so they add an electron that is capable of conduction in the conduction band. Such Group V atoms are therefore called “donors”. They are also known as “*n*-dopants” (with the “*n*” coming from “negative” because of the negative charge of the additional electron), and the resulting material can be called “*n*-type” or “*n*-doped”.

Acceptor atoms are also known as “*p*-dopants” (with the “*p*” coming from “positive” because of the effective positive charge of the additional hole), and the resulting material can be called “*p*-type” or “*p*-doped”.

Group III atoms, which include boron and gallium, have one fewer electrons in their outer “shell” compared to Group IV elements and those can therefore effectively subtract an electron from the valence band. These are known as “acceptors” (because they “accept” an electron from the valence band).

Once we add a moderate number of acceptor atoms, we see that the valence band is now only nearly full of electrons. In electron terms, we can now have conduction because there are states that electrons can move into. We can usefully give a name to such “absences of electrons”; we call them “holes”. It can now be much more convenient to describe the conduction in terms of the movement of these few “holes” rather than the movement of the very large number of electrons in this nearly-full band. Since these holes are describing the absence of an electron, it can be convenient to view them as having a positive charge. If we add a hole into the valence band, we have subtracted one electron’s worth of charge from the valence band, which is the same as adding a positive charge of the same magnitude.

Holes can be viewed as being the equivalent of bubbles in a liquid. Bubbles tend to rise in a liquid (which actually corresponds to the liquid falling), so to understand their potential energy, we have to “stand on our heads”. Lower hole energy corresponds to holes rising in our (electron energy) band diagrams, and holes will tend to gather at or near the top of the valence band, just as electrons in the conduction band would tend to gather at or near the bottom of the conduction band.

A “current” of bubbles going in one direction in an otherwise stationary liquid means that an amount of liquid of volume equal to the volume of the bubbles has to move in the opposite direction⁵⁷. For a variety of reasons, many of which are more subtle than they might at first appear, we can in practice usually view holes in the valence band as being positively charged particles of positive mass, and we will then usually get the right answers for the direction of the electrical current. (We return to this point in Chapter 6.)

⁵⁷ If you have dug a hole in the ground, but then decide it is in the wrong place, and it needs to be moved to the right, then to do so, you end up moving soil to the left.

Under normal conditions of room temperatures and moderate levels of doping, the simple model of one added electron in the conduction band for each added donor atom or one added hole in the valence band for each added acceptor atom is roughly correct in typical semiconductors. It is not, however, quite exactly correct, and that simple picture does not explain what happens as we change temperature or if we use a high doping density (a large number of dopant atoms per unit volume); it also does not explain why adding dopants to insulators, or even in some cases to wide band gap semiconductors, often has little or even sometimes no influence on their conductivity.

A more sophisticated model of doping is to say that adding in, say, a donor atom adds in an electron and also a “donor” energy level that the electron can occupy, as sketched in the left of Fig. 4.25. Under the normal conditions (e.g., room temperature or above) at which we operate semiconductor devices, most of the electrons associated with such donor atoms are effectively “ionized” by the available thermal energy to become electrons in the conduction band. Similarly, for acceptors and holes, most of the holes associated with the acceptor atoms are effectively “ionized” to become holes in the valence band.

Analogously to Fig. 4.24, we can also show these n-doped and p-doped semiconductors with ordinary position on the horizontal axis (Fig. 4.26). Again, this kind of picture is common when discussing device structures that are made with layers of different materials.

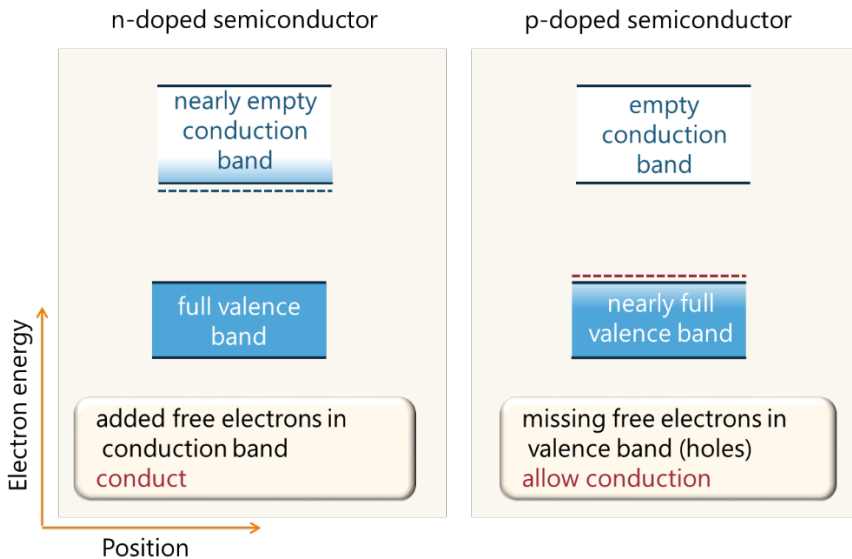


Fig. 4.26. Schematic illustration of bands in real space, in one direction, for an n-doped and a p-doped semiconductor.

In insulators, for a variety of reasons, adding impurities may not lead to them conducting electricity. One reason can be that the energies of donor and acceptor levels can be too far from the corresponding band edges, and the thermal ionization of those “dopant” atoms does not occur to a sufficient degree.

Quiz questions

- Q4.16.1 A band that is full of electrons conducts electricity well. True or False?
- Q4.16.2 A pure semiconductor material does not conduct electricity. True or False?
- Q4.16.3 Adding “holes” to a semiconductor – for example, by p -doping it - will necessarily reduce its electrical conductivity because it reduces the number of electrons. True or False?
- Q4.16.4 Metals only conduct electricity because they are impure materials, e.g., such as alloys. True or False?
- Q4.16.5 We should expect that if we add a small number of arsenic atoms to silicon, these arsenic atoms will behave as n -type dopants, effectively adding electrons into the conduction band. True or False?
- Q4.16.6 We should expect phosphorous would be a good p -type dopant in silicon. True or False?

Problems

- 4.16.1 At very low temperature in a metal, there will be no electrons in the conduction band. True or false? (Justify your answer briefly.)
- 4.16.2 A perfectly pure semiconductor material will have equal numbers of electrons in the conduction band and holes in the valence band, and so it will not conduct electricity because the electron and hole currents will exactly cancel. True or false? (Justify your answer briefly.)

4.17 Conclusions

Here we have seen that, starting from the Schrödinger equation, and with the addition of the key ideas of the Pauli exclusion principle and spin, we have been able to build up a deep framework for understanding atoms and materials, especially crystals.

The solution of the hydrogen atom problem lays the basis for understanding the orbitals of atoms. The Pauli exclusion principle, together with the approximate notion of each electron seeing an effective average potential from all the other electrons and atomic nuclei, lets us understand more complicated atoms. A simple picture of coupled systems gives us the basis for understanding core aspects of chemical bonding, such as bonding and antibonding states. Extending that picture into periodic systems, especially with the simplifying mathematical approach of the Bloch theorem, lets us see the emergence of bands of states in crystals. This band picture gives us a solid basis for understanding many of the electronic and optical properties of important materials for devices.

Though the details of all of these phenomena can become quite involved and difficult to calculate in practice, we now have the basic quantum mechanical ideas and vocabulary we need for use in a broad range of applications.

Chapter 5

Thermal distributions

5.1 Statistics and physics

Until the last 30 years of the 19th century, heat and the phenomena connected with it were generally regarded as being based on continuous variables. For example, energy was not presumed in any way to be discretized into “chunks”, nor were atoms or other particles considered only to occupy specific countable states. The concept of entropy was emerging from this background as scientists and engineers were trying to understand the limits to efficiency in steam engines. Entropy in such purely thermodynamic discussions is a very abstract idea and is quite difficult to understand. In turn, that makes the Second Law of Thermodynamics – which in one form states, essentially, that entropy always increases – even more difficult to comprehend.

The idea that thermodynamics, and entropy in particular, had anything to do with statistical distributions was initially an unusual concept. Indeed, there was at that time in the latter part of the 19th century no basis for saying that any kind of discrete states were involved in describing matter. Such ideas did not emerge in their own right until the advent of quantum mechanics in the 20th century. Despite this absence of the ideas of countable states for matter, these ideas of statistical distributions and a resulting firm underpinning for all of thermodynamics emerged from the work of James Clerk Maxwell, Josiah Gibbs, and, especially, Ludwig Boltzmann, primarily in the last part of the 19th century. These ideas remained very controversial throughout Boltzmann’s lifetime, only finally being vindicated convincingly in the early 20th century.

Arguably, the statistical basis behind all of thermodynamics makes it much more directly comprehensible. In particular, entropy then has a relatively straightforward and tangible meaning¹.

¹ The ideas of entropy are also closely tied in with the modern concept of information, though we will not be able to discuss that link in any detail here. “Bits” are, however, actually a measure or unit of entropy in information.

5.2 Tossing coins

We can start our discussion with a simple example of statistics with which we are all familiar – tossing coins. Suppose we toss a coin 4 times. We will get some set of “heads” and/or “tails” results. All the possibilities are shown in Fig. 5.1. Here we have used an up-arrow (“↑”) to indicate “heads” and a down-arrow (“↓”) to indicate “tails”. There are 16 possibilities altogether.

Microstate	Macrostate	Multiplicity	Excess, $2s_e$
↑↑↑↑	4 up (“heads”)	1	4
↑↑↑↓ ↑↑↓↑ ↑↓↑↑ ↓↑↑↑	3 up (“heads”) 1 down (“tails”)	4	2
↑↑↓↓ ↑↓↑↓ ↓↑↑↓ ↓↑↓↑ ↓↑↓↑ ↓↑↑↓	2 up (“heads”) 2 down (“tails”)	6	0
↓↓↓↑ ↓↓↑↓ ↓↑↓↓ ↑↓↓↓	1 up (“heads”) 3 down (“tails”)	4	-2
↓↓↓↓	4 down (“tails”)	1	-4

Fig. 5.1. Different possible results (“microstates”) for tossing a coin 4 times, grouped into “macrostates” depending on the number of ways of getting given numbers of “heads” and “tails”, with “multiplicities” representing the number of microstates in a macrostate. The “excess”, $2s_e$, is the number of “heads” minus the number of “tails” in a given macrostate.

We can call each possible set of four outcomes a “microstate”. Note that we can group the “microstates” into what we call “macrostates”. Each macrostate contains all the microstates that correspond to a particular number of “heads” and “tails”. The number of microstates in a macrostate is called the “multiplicity” of the macrostate.

Fig. 5.2 shows these multiplicities for this case of $N = 4$ coin tossings (top left graph), and for larger N , plotted against the “excess”, $2s_e$, which is the difference between the

number of “heads” and the number of “tails”². s_e itself is the number by which the number of “heads” exceeds the average number of “heads”, that average being $(N/2) = 2$ here for $N = 4$.

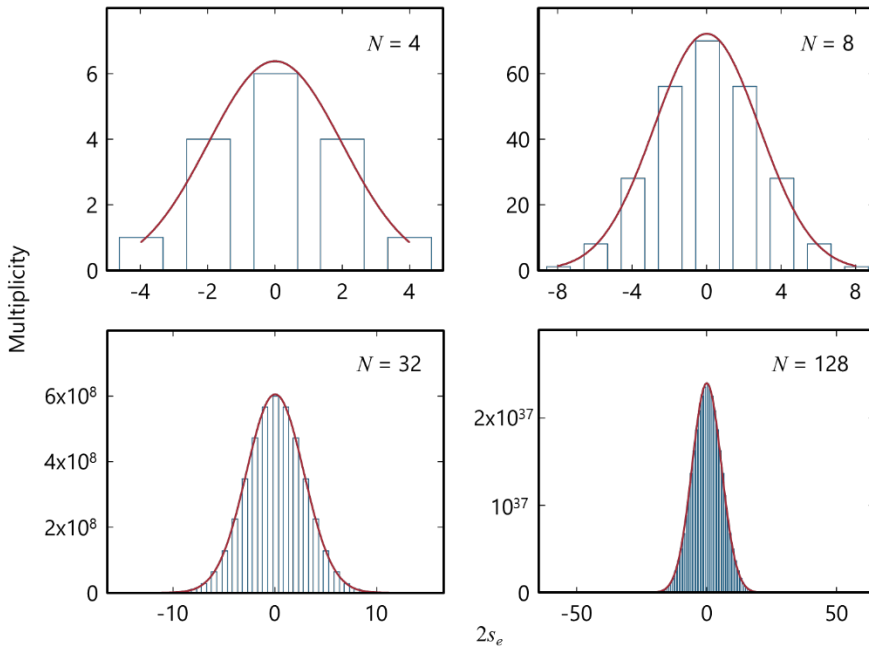


Fig. 5.2. Multiplicities (bars) for different N (number of coin tosses or number of spins) as a function of $2s_e$, the excess number of heads over tails or “spin-up” over “spin-down” values. The curves are the Gaussian approximation to the multiplicity for large N .

We can see various behaviors in Fig. 5.2. First, we note that the single most likely outcome is that we will have equal numbers of heads and tails, with 6 microstates corresponding to that macrostate – that is, with a multiplicity of 6. Outcomes or macrostates with differing numbers of heads and tails are less likely – that is, they have smaller multiplicities. We see that as we increase N , the number of times we toss the coin, this behavior persists. In all cases, as the difference between the number of heads and tails becomes larger – that is, as we move further from the center in any of the graphs in Fig. 5.2 – the multiplicity drops.

Second, we see that the multiplicity at the peak – that is, at $s_e = 0$ – becomes very large as we increase N . For large N , the number of sequences of coin tossings in which the number of heads and the number of tails is the same or similar becomes extremely large, and starts to dominate over all other possibilities³.

² Our notation here anticipates calling $2s_e$ the “spin excess” – the difference between the numbers of “spin-up” and “spin-down” in a set of spins.

³ This phenomenon is the source of the error in the “gambler’s fallacy”. The gambler thinks that just because the coin has come up “heads” several times in a row, it is more likely on the next coin toss that it will come up “tails”. The gambler believes this because he knows that, on the

Third, with increasing N , the distribution is getting wider – for $N = 128$ we can see by eye that the width of the distribution with s_e is maybe something like ~ 20 or so, which is obviously wider than in any of the other graphs in Fig. 5.2 – but, importantly, compared to the number N of coin tossings, it is becoming a smaller fraction.

Binomial distribution 16b

The formula that gives these bar graphs is quite straightforward, and is an example of the binomial distribution. To understand this counting, we can consider we have N coins in a row; we can think of them as all being “heads” to start with if we like, though it does not matter in the end for this argument⁴. We take some number k of them, and flip them over (to “tails”, if we like). We thereby create two sets, the flipped-over set with k coins and the un-flipped-over set with $N - k$. There are N ways we can choose the first coin to flip over, $N - 1$ ways we can choose the second, and so, down to $N - k + 1$ for our k th choice; multiplying these together gives us

$$N(N-1)(N-2)\dots(N-k+1) = \frac{N!}{(N-k)!} \quad (5.1)$$

There are also $k!$ different orderings in which we could have chosen which coins to flip over while still leaving us the same sets of flipped-over and un-flipped-over coins in the row, so we divide by $k!$, giving our final expression for the total number of different-looking rows of coins in which k of them are flipped over, that is

$$g = \frac{N!}{(N-k)!k!} \quad (5.2)$$

which is therefore our counting of the multiplicity of a macrostate corresponding to k “heads” and $N - k$ “tails”⁵.

average, the fraction of heads and tails does come out to about 50% each, and so he thinks there must be some mechanism, which he sometimes calls “the law of averages” (a non-existent statistical “law”), that somehow balances everything out. In fact, all that is going on is that there are just many more possible sequences in which there are approximately equal numbers of heads (H) and tails (T), and any sequence is presumed to be equally likely – HHHHHHHH is just as likely as HTTHTHH. This phenomenon of the dominance of sequences with near equal numbers of heads and tails is correctly viewed as an example of the “law of large numbers”, a statistical law that does exist.

⁴ Rather than talking about flipping coins over, we can just talk choosing to set some coins to tails, and setting all the others to heads.

⁵ The binomial coefficient $g = \frac{N!}{(N-k)!k!}$ is a standard result in combinatorics, sometimes

known as “ N choose k ”, and is also written in notations like $C(N, k)$, C_k^N and $\binom{N}{k}$.

For our arguments here, we prefer to write this expression in terms of s_e , which would give us $(N/2) + s_e$ for the number of “heads” and $(N/2) - s_e$ for the number of “tails”⁶. In that case, we will write for the multiplicity in this problem

$$g(N, s_e) = \frac{N!}{\left(\frac{N}{2} + s_e\right)! \left(\frac{N}{2} - s_e\right)!} \quad (5.3)$$

It is easy to check these formulas Eqs. (5.2) and (5.3) against the results in Fig. 5.1 and Fig. 5.2. For example, for $k = 1$ “heads” out of $N = 4$ coin tossings, we would have, from Eq. (5.2)

$$g = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times 1} = 4 \quad (5.4)$$

and for $s_e = 2$, an “excess” $2s_e = 4$, which corresponds to 4 more “heads” than “tails”, we would have from Eq. (5.3)

$$g(4, 2) = \frac{4!}{(2 + 2)! 0!} = \frac{4!}{4! 0!} = 1 \quad (5.5)$$

where we remember that $0! = 1$.

We want to be able to draw conclusions about large physical systems, with numbers of atoms that may be 10^{20} or more. To look at what happens as we increase to even larger N , we can use an approximation to this binomial distribution, which in turn is based on an approximation for the factorial $n!$ for large n . This approximation

$$\log(n!) \approx \frac{1}{2} \log 2\pi n + n \log n - n \quad (5.6)$$

is known as Stirling’s formula^{7,8}.

Note, incidentally, that in fundamental mathematical and physical expressions such as Eq. (5.6), logarithms are usually implicitly or explicitly to the base $e \approx 2.718281828459\dots$. Sometimes such “natural” logarithms are written as “ $\log_e$ ” or as “ $\ln$ ”, which is more explicit than just writing “ $\log$ ”. The “ $\ln$ ” (i.e., “el en”) notation is both explicit and compact, but unfortunately it is easily confused in most typefaces with “ $\ln$ ” or even “one n”. Because nearly all the logarithms we use will be such natural logarithms, we will use the notation “ $\log$ ” to refer to the natural logarithms; we will

⁶ We are implicitly assuming here for simplicity that N is even. This is not a necessary restriction for dealing with this problem, but it makes the algebra somewhat easier, and since N will be large in the following argument, this will not matter in the end.

⁷ Equivalently, $n! \approx \sqrt{2\pi n} (n/e)^n$ where e here is the base of the natural logarithms. Often, Stirling’s approximation is written dropping the first term as being relatively insignificant at large n , giving $\log(n!) \approx n \log n - n$, though we need the fuller form here for one further result.

⁸ James Stirling (1692-1770) was a Scottish mathematician.

explicitly use notations of the form “ $\log_2$ ” or “ $\log_{10}$ ” to refer to logarithms to the base 2 or to the base 10 if we need them.

Using this approximation, Eq. (5.6), we can derive a convenient result. For large N , we can approximately rewrite the expression in Eq. (5.3) as

$$g(N, s_e) \approx g(N, 0) \exp\left(-\frac{2s_e^2}{N}\right) \quad (5.7)$$

where

$$g(N, 0) \approx \sqrt{\frac{2}{\pi N}} 2^N \quad (5.8)$$

The proof of this result is relatively straightforward, starting from Eq. (5.3) and from Stirling’s formula, Eq. (5.6), and is left as an exercise for the reader⁹. A function of this type as in Eq. (5.7) is called a Gaussian, and it gives a characteristic “bell-shaped” curve.

The expression Eq. (5.8) is used to draw the limiting curves in Fig. 5.2. We can see by eye that, even for these relatively small values of N , this expression is a good approximation. We can now use Eq. (5.8) to examine the multiplicity for very large N . The relative multiplicity¹⁰ is plotted in Fig. 5.3 for N of 10^3 , 10^4 , 10^5 and 10^6 . As N becomes large, the absolute width does continue to grow, but the relative width becomes smaller, as we see in Fig. 5.3.

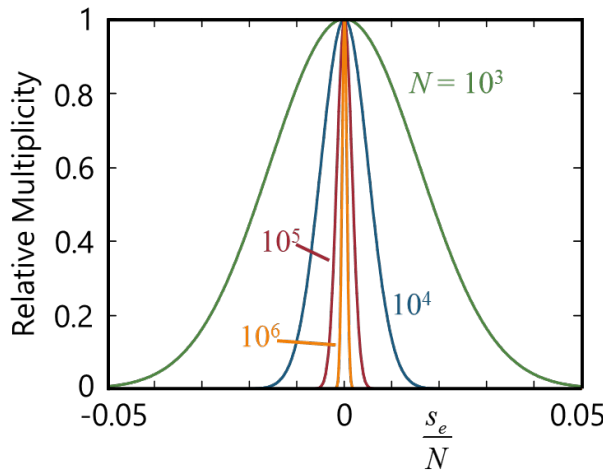


Fig. 5.3. Multiplicity relative to the maximum using the Gaussian approximation (i.e., plotting $\exp(-2s_e^2/N)$) for large N ,) as a function of s_e/N , the relative difference of the number of “heads” or “spin-up” states compared to the average.

⁹ Other than standard manipulations of logarithms, the only additional step required is approximating $\log(1+x) \approx x - x^2/2$ for small x .

¹⁰ By relative multiplicity we mean the multiplicity relative to that at $s_e = 0$.

We can see directly from Eq. (5.8) that the value of s_e for which the distribution falls to $1/e^2$ of its peak value is

$$s_{ec} = \sqrt{N} \quad (5.9)$$

which is therefore a characteristic width of the distribution. So, though the width of the distribution grows as $\sqrt{N}$, for the relative width we have

$$\frac{s_{ec}}{N} = \frac{1}{\sqrt{N}} \quad (5.10)$$

which expresses the fact that the relative width of the distribution decreases as the square root as N increases. So, for example, if we tossed a coin $N = 10^{20}$ times, the average number of heads or tails would be 0.5 within roughly 1 part in 10^{10} .

Note that nearly all the microstates are found in or very close (in terms of s_e / N) to the most likely macrostate, especially for large N . This is a quite general behavior for such statistical systems with large numbers of elements like this.

Quiz questions

Q5.2.1 If I toss a coin two times

- a) what is the total number of possible microstates (i.e., each corresponding to different sequences of heads and tails): 1, 2, 3, or 4?
- b) how many of these microstates are in the most probable macrostate: 0, 1, 2, or 3?

Q5.2.2 If I toss a coin 5 times, what is the total number of possible microstates (i.e., each corresponding to different sequences of heads and tails): 5, 8, 16, or 32?

Q5.2.3 If I toss a coin 8 times, what is the total number of possible microstates (i.e., each corresponding to different sequences of heads and tails): 128, 256, 512, 1024?

Q5.2.4 Suppose I toss a coin 8 times.

- a) How many microstates correspond to 4 heads and 4 tails: 16, 28, 32, or 70?
- b) How many microstates correspond to 6 heads and 2 tails: 4, 28, 32, or 70?
- c) How many microstates are in the most probable macrostate: 28, 56, 70, or 100?

Q5.2.5 If we use Stirling's approximation to calculate the multiplicities of macrostates, and we toss a coin 8 times, which of the following is closest to the answer given by Stirling's approximation in each case?

- a) The number of microstates in the most probable macrostate: 68.1, 70, 72.22 or 256?
- b) The number of microstates corresponding to 3 heads and 5 tails: 56, 53.29, 57.93, or 256?

Q5.2.6 If we toss a coin 64 times, according to Stirling's approximation, by approximately what value of the excess number of heads (i.e., $2s_e$) will the multiplicity have fallen off from its peak value by a factor $1/e^2 : 1, 4\sqrt{2}, 8$, or 16?

Problems

- 5.2.1* Following Fig. 5.1, draw all the possible “microstates” for tossing a coin 6 times. Then group these microstates into “macrostates” that contain all the microstates with the same numbers of “heads” and “tails”, with “multiplicities” representing the number of microstates in a macrostate. Which macrostate is the most probable state?
- 5.2.2 Consider $N = 10, 20, 30$, and 40 coin tossings. For $N = 10$ and 20, calculate any binomial coefficients by hand. For $N = 30$ and 40, you may have automatic calculators or some other resources do the job for you.
- For each N , calculate the multiplicity of the most probable macrostate using the binomial distribution formula and then again with Stirling's approximation. Calculate the percentage error of using Stirling's approximation in each case. How do the percentage errors change as N gets larger?
 - For each N , find the excess number of “heads” ($2s_e$) where the multiplicity of this macrostate falls to $1/e$ of the maximum multiplicity. How does the ratio s_e/N for this s_e change when N gets larger? [You may use Stirling's formula for this part of the question.]
- 5.2.3* If I toss a coin 100 times,
- calculate how many possible sequences of results there are in which only two of the results are “heads” and the others are all “tails”, and
 - estimate approximately how many possible sequences there are for which the number of “heads” and the number of “tails” are equal
- 5.2.4 The stadium at McRowe State University seats 50,420 and has an attendance of 43,340 at a game.
- Assuming that half of the attendees are McRowe State fans and can sit only in the north part of the stadium (which contains half of the seats), while the other half of the attendees can only sit in the south part, write out an expression for the number of different ways the seats can be occupied. (All fans are presumed created equal, so it does not matter who is sitting at an exact spot within their appropriate half of the stadium, only which seats are occupied). Note: because of the large numbers, use Stirling's formula to approximate any factorials in your expression.
 - Now assume there is no segregation between the two halves of the stadium, so any fan can occupy any seat. By approximately what factor is the number of possible seating arrangements smaller or larger than in the first case?

5.2.5 Using Stirling's approximation and the power series expansion of $\log(1+x)$ (up to second order in x) show that the binomial coefficient

$$g(N, s_e) = \frac{N!}{\left(\frac{N}{2} + s_e\right)! \left(\frac{N}{2} - s_e\right)!}$$

for large N can be written approximately as

$$g(N, s_e) \approx g(N, 0) \exp\left(-\frac{2s_e^2}{N}\right) \text{ where } g(N, 0) = \sqrt{\frac{2}{\pi N}} 2^N$$

5.3 Two-state spin systems 16c

So far, we have presumed that every such outcome of heads and tails is possible, as indeed it is for tossing a coin. A related physical system would be a set of spins, such as some set of electrons with negligible interaction between different ones¹¹. Just as each coin could be either heads or tails, so also each spin (or electron) could be “spin-up” (“↑”) or “spin-down” (“↓”). If there are no other constraints on spins being up or down, then we obtain exactly the same kind of counting for spins as for the coins. Indeed, anticipating this, we used an up-arrow (“↑”) as the analog of “heads” and a down-arrow (“↓”) as the analog of “tails” in Fig. 5.1.

In real physical systems, however, different possible microstates may have different energies. If for some reason there is a fixed energy available or there is some other constraint on energy, it may be that not all the microstates we could write down would be possible or what we call “accessible”. We should define that concept.

Given whatever physical constraints there are on the system, an accessible microstate is one that is physically possible.

Now we make an important physical postulate, one that underlies our entire statistical analysis here of physical systems¹².

¹¹ We want this “negligible interaction” so we can talk meaningfully about individual spins without having to consider coupled spin systems whose energies might be influenced by that coupling.

¹² Essentially here, we are dividing possible microstates into ones that are “accessible” and ones that are not. For many physical situations, this is a very good approximation. We can imagine situations in which it is not, however. If some barrier between two gasses is very slightly permeable, then we cannot quite neglect the possibility that an atom or molecule from one side will get to the other side. Then, such a microstate is a possibility, but it is less likely than many other possibilities, at least if we only wait for some finite time to declare “equilibrium”. We can extend statistical mechanics with probabilistic weightings to cover such situations, though we will not do that here. That extension is, however, quite common in the idea of entropy as used in information theory. For example, some letters in an alphabet are more likely to occur than others in a written message. “E” is a particularly common letter in English, for example, whereas

For a physical system in equilibrium, each accessible microstate is equally probable.

By “equilibrium” here, we mean essentially, after everything has settled down and appears not to be changing any more, at least as we observe it macroscopically. Of course it is possible that we start the system in some very specific microstate, so the probability of that one dominates over all the others at that starting “non-equilibrium” condition. But once we have left the system to equilibrate, on measuring the system, any accessible microstate is presumed to be just as likely as any other.

Now we will consider this system of spins, and use it as a model to understand what happens in real physical systems. We are not going to attempt general proofs for physical systems more widely; we merely presume that the behaviors we can deduce from this simple kind of system can be generalized to others.

Note, incidentally, that we are presuming that there is some physical process that allows the spins to “flip” from one state to the other (i.e., from “spin-down” to “spin-up” or *vice versa*) – possibly as a result of some collisions, for example - so that all possible microstates of spin-up and spin-down for each spin are “accessible” in this sense.

To model physical systems more realistically, we need to allow the microstates to have energies that may not all be the same. For our spin system, we can imagine that we apply a magnetic field of magnitude B along the “up-down” axis of the spins. For a small magnet of magnetic moment (magnet “strength”) μ_s , the energy of a “spin-up” “magnet” in the magnetic field can be written as

$$E_\mu = -\mu_s B \quad (5.11)$$

A spin with “spin-down” will have exactly the opposite energy in the magnetic field. To the extent that some spins are “up” and others are “down”, the energies of equal numbers of “up” and “down” spins will cancel out when we work out the total energy, U . Hence, the total energy only depends on the “spin excess” $2s_e$, specifically being¹³

$$U(s_e) = -2\mu_s B s_e \quad (5.12)$$

We are interested for the moment in some closed physical system in which this total energy U is fixed for some reason, which means we are presuming conservation of energy for this set of spins in what we will call a “closed” system.

Systems in thermal contact

Now let us presume we have two systems, as shown in Fig. 5.4, each containing such spins. Initially, we imagine these systems are separate, and each is well insulated from its environment. System 1 has a total energy U_1 , a number of “spins” N_1 , and a “spin excess” $2s_{e1}$, with corresponding quantities U_2 , N_2 , and $2s_{e2}$ for system 2.

“q” is not, so entropy in information theory when applied to written messages has to account for these probabilities.

¹³ Remember that $2s_e$ is the number by which we have more “spin-up” than “spin-down”; it should not be confused with the spin quantum number s .

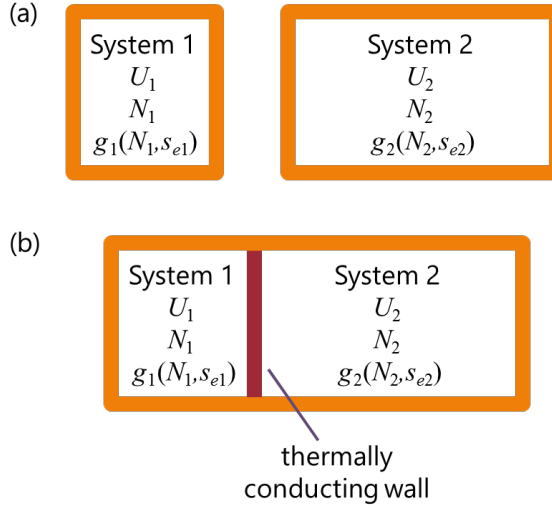


Fig. 5.4. Two systems containing spins, (a) initially separate (top), and (b) later joined through a thermally conducting wall (bottom)

We presume that the total energy in the combined system is to be conserved, even when we join the systems through some thermally conducting wall. So, at all times, the sum

$$s_e = s_{e1} + s_{e2} \quad (5.13)$$

which will give the total energy of the system through Eq. (5.12), will be fixed or “conserved” even after we join the systems. (This thermally conducting wall conducts heat, but does not allow spins to move through it, so the numbers N_1 and N_2 remain fixed throughout.)

Our goal now is to deduce what will be the most likely possible macrostate for splitting the total s_e between the two systems when we have reached thermal equilibrium after joining the two systems through the thermally conducting wall. Since we know that, for large systems, the most likely macrostate and those close to it dominate the possible microstates, if we know this most likely macrostate and its properties, we will essentially know the properties of the system at thermal equilibrium.

Presuming N_1 and N_2 are both large, for a given s_{e1} and N_1 in system 1, we know the multiplicity of the macrostate for system 1 from Eq. (5.7), which is

$$g_1(N_1, s_{e1}) \simeq g(N_1, 0) \exp\left(-\frac{2s_{e1}^2}{N_1}\right) \quad (5.14)$$

and similarly for the multiplicity for system 2, $g_2(N_2, s_{e2})$.

The multiplicity of the macrostate of the entire system is simply the product of the multiplicities of the parts; in given macrostates of systems 1 and 2, if there are g_2 microstates of system 2 possible for each microstate of system 1, and there are g_1 possible microstates of system 1, then the total number of possible microstates for the entire system is $g_1 g_2$. So, if there are $g(N_1, s_{e1})$ possible microstates for system 1 and $g(N_2, s_{e2})$ for system 2, then for the entire system, there are

$$g_{tot} \approx g(N_1, 0)g(N_2, 0)\exp\left(-\frac{2s_{e1}^2}{N_1}\right)\exp\left(-\frac{2s_{e2}^2}{N_2}\right) \quad (5.15)$$

possible microstates for these choices of s_{e1} and s_{e2} that characterize the macrostate of systems 1 and 2, respectively. Our task now is to find out how we should split up s_e between s_{e1} and s_{e2} so as to maximize the multiplicity g_{tot} .

Since s_e itself is fixed, then we can rewrite Eq. (5.15) in terms only of a fixed s_e and a variable s_{e1}

$$g_{tot}(s_{e1}) \approx g(N_1, 0)g(N_2, 0)\exp\left(-\frac{2s_{e1}^2}{N_1}\right)\exp\left(-\frac{2(s_e - s_{e1})^2}{N_2}\right) \quad (5.16)$$

Maximizing multiplicities for two-state spin systems

Now we can differentiate Eq. (5.16) with respect to s_{e1} , setting the result to zero to find the choice of s_{e1} that maximizes g_{tot} . It is slightly easier to do this if we take the logarithm and find the maximum of the logarithm (if a positive function is at a maximum, so also is its logarithm). Taking the logarithm of both sides of Eq. (5.16) gives

$$\log[g_{tot}(s_{e1})] \approx \log[g(N_1, 0)g(N_2, 0)] - \frac{2s_{e1}^2}{N_1} - \frac{2(s_e - s_{e1})^2}{N_2} \quad (5.17)$$

Taking the derivative and setting this to zero gives

$$\frac{d \log[g_{tot}(s_{e1})]}{ds_{e1}} \approx -\frac{4s_{e1}}{N_1} + \frac{4(s_e - s_{e1})}{N_2} = 4\left(\frac{s_{e2}}{N_2} - \frac{s_{e1}}{N_1}\right) = 0 \quad (5.18)$$

The second derivative – that is, $-4[(1/N_1) + (1/N_2)]$ – is negative, which confirms this is a maximum. Hence we conclude that, to maximize the multiplicity of the combined system,

$$\frac{s_{e2}}{N_2} = \frac{s_{e1}}{N_1} \quad (5.19)$$

Since in a given magnetic field the energy of a given one of the systems is just correspondingly proportional to s_{e1} or s_{e2} , that is, from Eq. (5.12)

$$U_1(s_{e1}) = -2\mu_s B s_{e1} \quad \text{and} \quad U_2(s_{e2}) = -2\mu_s B s_{e2} \quad (5.20)$$

the statement Eq. (5.19) is equivalent to the statement that the energy per spin in the two systems is the same – that is, from Eqs. (5.19) and (5.20)

$$\frac{U_1}{N_1} = \frac{U_2}{N_2} \quad (5.21)$$

Of course, if the energy per spin is the same on each side, then the energy per spin in the entire system is also the same. That is, with

$$U = U_1 + U_2 \quad (5.22)$$

and

$$N = N_1 + N_2 \quad (5.23)$$

then

$$\frac{U_1}{N_1} = \frac{U_2}{N_2} = \frac{U}{N} \quad (5.24)$$

and similarly¹⁴

$$\frac{s_{e2}}{N_2} = \frac{s_{e1}}{N_1} = \frac{s_e}{N} \quad (5.25)$$

So, by example here with this simple system, we have uncovered a relatively straightforward result: the average energy for each similar microscopic system – here the spins – is the same for systems at thermal equilibrium. To understand the deeper meaning of this, we need to generalize some more, and introduce two concepts that are new to our discussion here – entropy and temperature.

Quiz questions

- Q5.3.1 Suppose I have a set of spins, with a fixed positive number more that are spin up rather than spin down. I apply a magnetic field B along the “up-down” axis as in the lecture, and we note that the magnetic moment s a positive number. The energy of the system decreases as a result of applying the field. True or False?
- Q5.3.2 Consider two (sub)systems, each of which is in a macrostate with some specific multiplicity. In considering the entire system of these two (sub)systems combined, the multiplicity of the macrostate of the entire system is the sum of the multiplicities of the two (sub)systems. True or False?
- Q5.3.3 Consider two systems with spins, each with some excess number of spin-up over spin down, and in the presence of a magnetic field that is the same for both systems. We join the systems with a thermally conducting wall to make a combined system, but the systems are otherwise well insulated from the environment. Choose the correct statement.
- a) The overall energy of the combined system will decrease after they are joined.

¹⁴ Though this result is perhaps obvious from the physical discussion of equal energy per spin, we can prove it algebraically as well. From Eq. (5.19), we have $s_{e2} = N_2 s_{e1} / N_1$, so $s_e = s_{e1} + s_{e2} = s_{e1}[1 + (N_2 / N_1)] = s_{e1}[(N_1 + N_2) / N_1] = s_{e1}N / N_1$, hence $s_e / N = s_{e1} / N_1$.

- b) The energy of the combined system will stay the same after they are joined.
- c) Whether the energy of the combined system increases or decreases depends on the sign (i.e., direction, pointing up or pointing down) of the magnetic field.

Q5.3.4 We have two systems, consisting of a large fixed number of identical particles, each of which can be “spin-up” or “spin-down”, and in a uniform magnetic field along the “up-down” direction. The systems are then joined by a thermally conducting wall to make a combined system, though the systems are otherwise thermally insulated from the environment. After combining the two systems and allowing them to come to thermal equilibrium ...

- a) The total number of spin-up and spin-down particles remains the same. True or False?
- b) The average energy per spin necessarily settles to zero. True or False?
- c) The fraction of spin-up particles on both sides is (at least approximately) the same on the average. True or False?

Problems

5.3.1* Consider a system of 6 spins in the presence of a magnetic field B . The energy of a macrostate with spin excess $2s_e$ is given by $U(s_e) = -2\mu_s B s_e$. The system is at equilibrium. Presume for the purposes of this question that each possible microstate is equally likely, independent of its energy.

- (a) What is the probability of the system having energy $U = 4\mu_s B$?
- (b) Derive an expression for the expected value (i.e., the average value) of the energy. The expected value can be calculated by the formula $\langle X \rangle = \sum_n P(X_n) X_n$, where $P(X_n)$ represents the probability that the variable X takes value X_n .

5.3.2 We have two sets of “spins” (such as electrons), one on the “left” and one on the “right” that are initially in separate boxes. A magnetic field (the same in each box) is applied so that spin-up spins have a specific amount more energy than spin-down spins. The number of spins is fixed, with 20 spins in the “left” box, and 80 spins in the “right” box. Initially, in the box on the left, 15 spins are spin-up and 5 are spin-down, and in the box on the right, 25 are spin-up and 55 are spin-down. We then join the boxes through a thermally conducting wall that can pass energy (but not spins) between the boxes, but otherwise both boxes are isolated from the environment at all times. After we let the combined system come to thermal equilibrium, what is the most likely configuration? In other words, what is the most likely number of spins being spin-up in each box?

5.4 Entropy and temperature

We will continue to think about the two systems as in Fig. 5.4 (b), but now generalized beyond just the spin systems we were considering. We will presume as before N_1 and N_2 quantum “particles” of the same kinds (like the spins) on each side respectively; N_1 and N_2 will be fixed numbers with a sum as in Eq. (5.23) as before. We will also still consider energies, U_1 and U_2 , respectively, in the two systems, now without restricting the systems to having any specific form, but still with a sum as in Eq. (5.22).

We still presume, however, that we are dealing with specific energy states that are therefore countable, and hence can allow us to construct multiplicities associated with these given energies. Hence for specific energies¹⁵ U_1 and U_2 we will have multiplicities for the two systems of $g_1(N_1, U_1)$ and $g_2(N_2, U_2)$ respectively.

To obtain the total multiplicity $g(N, U)$ for the combined system, we should sum up the multiplicities for every possible combination of U_1 and U_2 that corresponds to the same total energy U as in Eq.(5.22). So we have

$$g(N, U) = \sum_{U_1} g_1(N_1, U_1) g_2(N_2, U - U_1) \quad (5.26)$$

where we are summing over all the possible values of the energy U_1 (for example, from 0 to U). Because U_1 and U_2 must add up to the total energy U , as given by Eq.(5.22), we have substituted $U - U_1$ for U_2 .

Formally, to evaluate the multiplicity we should perform this summation. We presume, however, that the multiplicity overall will be dominated by the term in the sum for which the product $g_1(N_1, U_1) g_2(N_2, U - U_1)$ is largest, or at least a group of terms close to that, so we will consider just such terms, finding the conditions under which they are maximized; we presume that doing so will effectively also maximize the sum as in Eq. (5.26).

In general, if we make small (actually infinitesimal) changes in quantities such as U_1 and U_2 , we will make small (actually infinitesimal) changes in the multiplicity g that depends on U_1 and U_2 . This small change is called the differential. We note, first, that, since g_1 does not depend on U_2 and g_2 does not depend on U_1 (at least when we are regarding U_1 and U_2 as independent variables for the moment), that

$$\frac{\partial(g_1 g_2)}{\partial U_1} = \left(\frac{\partial g_1}{\partial U_1} \right) g_2 \quad \text{and} \quad \frac{\partial(g_1 g_2)}{\partial U_2} = g_1 \left(\frac{\partial g_2}{\partial U_2} \right) \quad (5.27)$$

¹⁵ The restriction to absolutely specific values of energies is unrealistic and somewhat problematic – there might be only one possible microstate of the system associated with one “exact” energy. We can, however, imagine we are considering some small range of energies δU about any specific value U in considering our macrostates, hence avoiding this formal problem.

Hence the differential can be written for this specific case as¹⁶

$$dg = \left(\frac{\partial g_1}{\partial U_1} \right)_{N_1} g_2 dU_1 + g_1 \left(\frac{\partial g_2}{\partial U_2} \right)_{N_2} dU_2 \quad (5.28)$$

The idea of this differential is that it would formally tell us the small change in the multiplicity g that would result if we made small changes in U_1 and U_2 , presuming that we somehow knew the partial derivatives in Eq. (5.28).

In our particular situation here, where total energy is conserved, we know that if U_1 increases by some amount then U_2 must decrease by an equal amount, so

$$dU_1 = -dU_2 \quad (5.29)$$

If the system is to be in the macrostate with the largest multiplicity, then its multiplicity $g = g_1 g_2$ should be at a maximum as far as the choice of U_1 is concerned. Therefore, if we were to make an infinitesimally small change in U_1 , there should be no change in $g_1 g_2$. Hence we can write that the differential of $g = g_1 g_2$ should be zero, and using Eq. (5.29) we have

$$dg = \left(\frac{\partial g_1}{\partial U_1} \right)_{N_1} g_2 dU_1 - g_1 \left(\frac{\partial g_2}{\partial U_2} \right)_{N_2} dU_1 = 0 \quad (5.30)$$

from which we can conclude that

$$\frac{1}{g_1} \left(\frac{\partial g_1}{\partial U_1} \right)_{N_1} = \frac{1}{g_2} \left(\frac{\partial g_2}{\partial U_2} \right)_{N_2} \quad (5.31)$$

or, equivalently,

$$\left(\frac{\partial \log g_1}{\partial U_1} \right)_{N_1} = \left(\frac{\partial \log g_2}{\partial U_2} \right)_{N_2} \quad (5.32)$$

So, we are concluding that, in thermal equilibrium, these two partial derivatives should be equal. If these two are equal, we will have found the conditions for the product $g_1 g_2$

¹⁶ Remember that the notation for the partial derivative $\left(\frac{\partial g_1}{\partial U_1} \right)_{N_1}$ means the derivative with respect to U_1 with N_1 held constant, and is formally defined as

$$\left(\frac{\partial g_1}{\partial U_1} \right)_{N_1} = \lim_{\Delta U_1 \rightarrow 0} \frac{g_1(N_1, U_1 + \Delta U_1) - g_1(N_1, U_1)}{\Delta U_1}$$

With partial derivatives generally, we sometimes do not specify explicitly which variables are being held constant if that is presumed obvious. So, for this derivative here, we do not explicitly mention that U_2 is being held constant by listing it as a subscript beside N_1 , though we could do that if we felt it was important for clarity.

to be maximized, and we are arguing that, since such macrostates dominate the sum of the multiplicities over all possible macrostates (i.e., Eq. (5.26)), we will have found the conditions for maximizing that multiplicity overall.

Defining entropy and temperature 18a

Usefully now we can define a quantity, which we call the “entropy”¹⁷

$$\sigma(N, U) \equiv \log g(N, U) \quad (5.33)$$

The key idea of entropy is that, for some given macrostate, it is the log of the multiplicity.

Note that, though the multiplicity of some combined system is the product of the multiplicities of the individual systems, because of the use of the logarithm, the entropy of the combined system is the sum of the entropies¹⁸. So, if we have two systems, 1 and 2, in macrostates with multiplicities g_1 and g_2 respectively, and hence with entropies

$$\sigma_1 = \log g_1 \text{ and } \sigma_2 = \log g_2 \quad (5.34)$$

then the total entropy of the joint system will be

$$\sigma_{tot} = \log(g_1 g_2) = \log g_1 + \log g_2 = \sigma_1 + \sigma_2 \quad (5.35)$$

Continuing with our argument here, we have, from Eq. (5.32) and the definition Eq. (5.33), for the macrostate with the largest multiplicity (the most probable configuration)

$$\left(\frac{\partial \sigma_1}{\partial U_1} \right) \Big|_{N_1} = \left(\frac{\partial \sigma_2}{\partial U_2} \right) \Big|_{N_2} \quad (5.36)$$

This is the condition for thermal equilibrium for two systems in thermal contact. We could restate this in words as “the rate of change of entropy with energy is the same for all systems in thermal equilibrium with each other” (at least for the case here of fixed numbers of particles in each system).

We are used to the idea that when systems are in thermal equilibrium with one another that it is the temperature that is the same. That is, for two such systems, with temperatures T_1 and T_2 , respectively, in thermal equilibrium, we expect them to have the same temperature

$$T_1 = T_2 \quad (5.37)$$

¹⁷ We can think of this particular version as being a “statistical” definition of entropy, sometimes called “fundamental entropy”. Different ways of stating entropy can use logarithms to different bases (information theory uses logarithms to the base 2) or have constants multiplying them (thermodynamics typically multiplies by Boltzmann’s constant). Such minor differences of definition make no difference to the underlying concept of entropy; all share the core idea of the logarithm of the multiplicity.

¹⁸ Remember $\log xy = \log x + \log y$

Hence we can now identify the quantities in Eq. (5.36) as being some function of temperature. To accord with our conventional understanding of temperature and its prior use in thermodynamics, we use the reciprocal of these quantities as the temperature, and also use a constant to get the temperature into the unit with which we are familiar, that is

$$\frac{1}{T} = k_B \left(\frac{\partial \sigma}{\partial U} \right) \bigg|_N \quad (5.38)$$

where k_B is Boltzmann's constant

$$k_B \approx 1.380\,6488 \times 10^{-23} \text{ J K}^{-1} \quad (5.39)$$

The Boltzmann constant is only there because of our system of units. It can be more convenient to work with “fundamental temperature”, τ , which we can define as

$$\frac{1}{\tau} = \left(\frac{\partial \sigma}{\partial U} \right) \bigg|_N \quad (5.40)$$

or equivalently

$$\tau = k_B T \quad (5.41)$$

Note that the fundamental temperature has dimensions of energy, which is the real unit of temperature. In that sense, even the kelvin unit¹⁹ is redundant, though it can be practically more convenient²⁰. Various other temperature scales are also in use based on the kelvin unit or “degree” Celsius or the Fahrenheit degree (Fahrenheit, Rankine)²¹.

¹⁹ Temperatures in the modern SI units are given in kelvin (not technically “degrees kelvin”, so there is no need for the “°” symbol for degrees when using kelvin, though in practice this form is often used, either in words or with the degree symbol), and indicated by the letter K. The Kelvin scale is zero at the absolute zero of temperature. Temperature in degrees Celsius is temperature in K plus 273.15. Equivalently, absolute zero is -273.15 °C. Note, incidentally, that all the unit names in the SI units system (see Appendix A), when written as a word, start with lower case, as in kelvin. We use the upper case Kelvin when referring to the person or to the temperature scale.

²⁰ A temperature of 300K, often taken as a rough simple number to describe “room temperature”, is $\sim 4.14 \times 10^{-21} \text{ J}$, not a particularly convenient number for everyday use!

²¹ The Celsius scale is defined with degrees of the same magnitude as 1 K, but with a different starting point. Until 1948, this scale was known as “centigrade” before being renamed to honor its creator, Anders Celsius (though interestingly, he actually had defined it the other way round, with 0 as the boiling point of water and 100 as the freezing point). The modern definition of the Celsius scale is that the zero is at 273.15 K. In loose terms, the Celsius scale runs from the freezing point of water at 0 °C to the boiling point at 100 °C, though the formal definition is as given here in terms of kelvin. The Rankine scale is a thermodynamic temperature scale starting at absolute zero, just like kelvin, though the degrees are the same magnitude as those in the Fahrenheit scale. Absolute zero is at -459.67 °F. The zero in the Fahrenheit scale may come from the lowest temperature to which Daniel Fahrenheit (who invented the scale and made one of the first good thermometers) could cool brine. Loosely, the Fahrenheit scale is 32° F at the

In thermodynamics, we write

$$\frac{1}{T} = \left(\frac{\partial S}{\partial U} \right) \bigg|_N \quad (5.42)$$

where the (thermodynamic) entropy S corresponds with our statistical “fundamental” entropy through

$$S = k_B \sigma \quad (5.43)$$

Again, units for entropy are technically unnecessary since it is really a pure number, being the logarithm of a number, but for historical reasons relating to the emergence of entropy first as a concept in thermodynamics, that thermodynamic statement of entropy has units of J K^{-1} .

We can therefore write the thermodynamic version of entropy explicitly in terms of the logarithm of the multiplicity g , formally using Eqs. (5.33) and (5.43), giving

$$S = k_B \log g \quad (5.44)$$

This is perhaps the most important equation in the whole field of statistical physics²², and expresses the core meaning of the idea of entropy. It finally gives a tangible meaning to what was otherwise a very difficult abstract concept. In this statistical sense, entropy is a logarithmic measure of the uncertainty of which microstate the system is in²³.

freezing point of water (32° F is defined to be the same temperature as 0° C) and 212 °F at the boiling point (though with the modern definitions, the actual boiling point of water is not quite exactly that). That separation between the freezing and boiling points might seem arbitrary, but note that it is 180°. The notion of 180° in a scale was already known from the idea of 180° in the measurement of geometrical angles, and 180 is one of those numbers that easily divides by many other integers (e.g., 2, 3, 4, 5, 6, 9, 10, 12, 15, 18, 20, 30, 45, 60, 90), which was convenient in the era before modern calculators.

²² The version $S = k \log W$ is engraved on Boltzmann’s tombstone.

²³ We could also say, therefore, that entropy is a measure of ignorance! This is a confusing point in physics. The system itself is not confused about what microstate it is in – it is in some specific microstate. But in our example here, we seem to be assigning two different entropies to the same microstate (log 4 and log 6). However, it is not the microstate that has entropy, and in this sense entropy is not a property of the physical system. (Note, incidentally, that this is an interpretation that some physicists might dispute.)

Returning to the idea of our ignorance, if we know nothing about the system’s specific state other than the total energy, and there are 6 possible such microstates of the system, then the entropy is log 6. If we can distinguish different macrostates, say by looking at the magnetization in each part of the system, and if we find it in the macrostate corresponding to zero magnetization in each part of the system (so one spin up and one spin down in each part), then there are 4 such microstates, and, armed with this information, the entropy is log 4. Hence, we could say that entropy is actually a measure of our ignorance of what microstate the system is in, and that ignorance has been reduced, from log 6 to log 4, by our measurement of the magnetization in each part of the system.

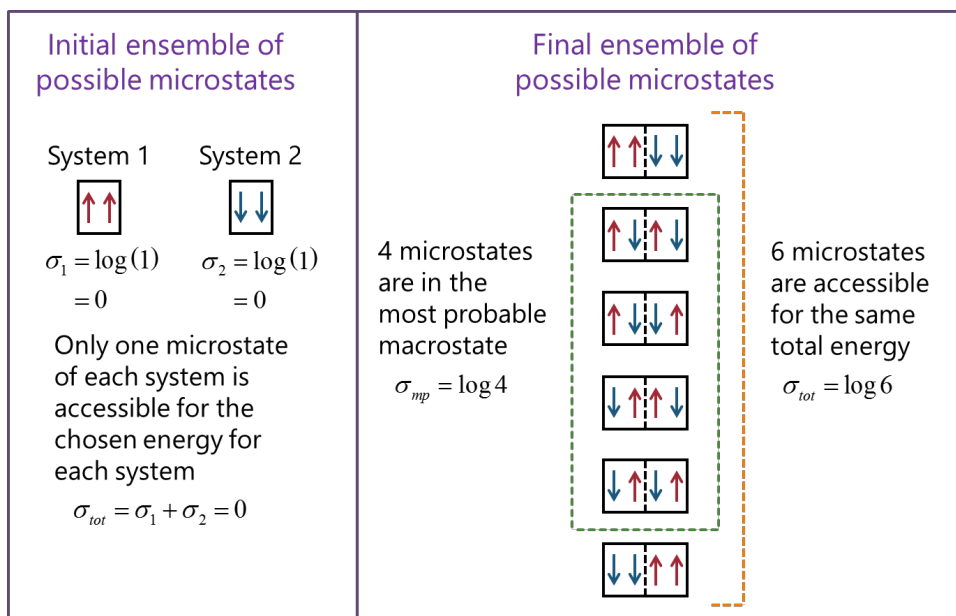


Fig. 5.5. Microstates and entropies for two systems, each consisting of two spins, before (left) and after (right) joining them.

We can look at an example of bringing two systems into thermal contact, in which energy can be exchanged between the two systems. Fig. 5.5 shows a very simple case. We have two systems, system 1 and system 2, with energies U_1 and U_2 respectively. Each one has two spins in it. Initially, system 1 has both spins “up”, and system 2 has both spins “down”. We presume that the same magnetic field B is applied to both systems, so the starting energies associated with these systems are, from Eq. (5.12), for system 1,

$$U_1 = -2\mu_s B \quad (5.45)$$

because we have two “spin-up” in system 1, and for system 2, which has two “spin-down”,

We often in practice regard entropy as being a property of the system because, for large systems, we can say with a very high degree of likelihood that the system will be in a microstate that is a member of the most probable macrostate or one very similar to it. Hence we can predict – in principle statistically, but in practice with a very high degree of certainty for even moderately large systems – what the properties of a system will be, for example after some exchange of heat. Because of this very high certainty for even moderately large systems, it can appear that entropy is a property of the system because it is a predictable outcome of some interaction, such as exchange of heat. So, in thermodynamics (which is essentially something we use just for large systems in this sense), entropy is treated as if it is a property of the system, and, at least in a practical sense, this works quite well.

For a large system, our brains are not large enough to hold the details of the actual microstate of some large number of atoms or molecules, so our ignorance in such situations is unavoidable, so we should not feel too bad about this!

$$U_2 = +2\mu_s B \quad (5.46)$$

There is only one starting microstate of each of systems 1 and 2 that corresponds to these two specific energies. For that reason, the starting entropy of each of system 1 and 2 is $\log 1 = 0$, and the total entropy σ_{tot} of the two systems is the sum of these, which also is mathematically 0.

After we bring the two systems together and allow them to exchange energy, there are 6 accessible microstates all corresponding to the same total energy. There are 4 of these in one macrostate, which is the most probable one; all the microstates in this macrostate have the same energy in each of the (sub) systems 1 and 2. Note that, even for this very simple example, 2/3 of the microstates are in the same macrostate. Nearly all the entropy of the combined system, $\sigma_{tot} = \log 6$, is associated with that most probable macrostate, for which the entropy is $\sigma_{mp} = \log 4$; explicitly, $(\log 4) / (\log 6) \approx 0.77$ – that is, $\sim 77\%$.

Entropy and heat flow 18b

Now that we have defined the notion of entropy and have formally defined temperature in terms of entropy and energy, we can examine the flow of heat between bodies of different temperatures.

We know that the total entropy of the two systems or “bodies” is just the sum of the individual entropies, as we wrote in Eq. (5.35). Note that the entropy σ_1 of system or “body” 1 does not depend on the energy U_2 of body 2, and similarly the entropy σ_2 of system or “body” 2 does not depend on the energy U_1 of body 1. So, if we changed the energy of body 1 by a small amount ΔU_1 and changed the energy of body 2 by some small amount ΔU_2 , then just by the normal properties of derivatives, we would have a total change of entropy of the combined system of

$$\Delta\sigma = \left(\frac{\partial\sigma_1}{\partial U_1} \right)_{N_1} \Delta U_1 + \left(\frac{\partial\sigma_2}{\partial U_2} \right)_{N_2} \Delta U_2 \quad (5.47)$$

Here we have formally written in these derivatives that the number of particles in each system is being held constant, which we presume is the case because we are only considering possible changes in energy here, not movement of particles from one body to the other.

Suppose we allow a small amount of heat ΔU to flow from body 1 to body 2. Now, since we are discussing the transfer of heat from one body to the other, if body 1 loses energy ΔU (so $\Delta U_1 = -\Delta U$) and body 2 gains energy ΔU (so $\Delta U_2 = \Delta U$), then Eq. (5.47) becomes

$$\Delta\sigma = \left(\frac{\partial\sigma_1}{\partial U_1} \right)_{N_1} (-\Delta U) + \left(\frac{\partial\sigma_2}{\partial U_2} \right)_{N_2} (\Delta U) = \left(-\frac{1}{\tau_1} + \frac{1}{\tau_2} \right) \Delta U \quad (5.48)$$

where we have used the definition Eq. (5.40) for temperature. In conventional thermodynamic notation, this result becomes, on multiplying both sides by Boltzmann’s constant,

$$\Delta S = \left(-\frac{1}{T_1} + \frac{1}{T_2} \right) \Delta U \quad (5.49)$$

So, if $T_1 > T_2$, so $(1/T_2) > (1/T_1)$, then transfer of a positive amount of energy or “heat” ΔU from body 1 to body 2 leads to an increase of entropy. The increase in entropy of the colder body is greater than the decrease in entropy of the hotter body. Explicitly, we could rewrite Eq. (5.49) as

$$\Delta S = \Delta S_1 + \Delta S_2 \quad (5.50)$$

where ΔS_1 and ΔS_2 are the respective changes in entropy of the two bodies, given by

$$\Delta S_1 = -\frac{\Delta U}{T_1} \quad \text{and} \quad \Delta S_2 = \frac{\Delta U}{T_2} \quad (5.51)$$

Here, then, the entropy of “hotter” body 1 has decreased with the flow of heat energy out of it, while the entropy of “colder” body 2 has increased with the flow of heat energy into it, with the direction of heat flow being such as to increase the entropy overall.

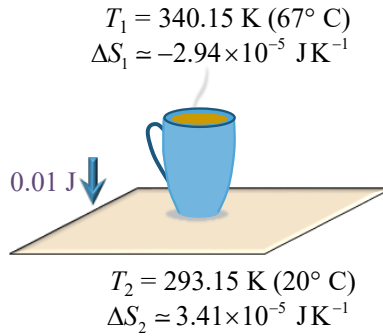


Fig. 5.6. Illustration of the transfer of heat from a hot cup of coffee to a room-temperature counter-top, showing the entropy changes for a transfer of 0.01 J of heat.

It is instructive to understand the magnitude of what is happening here for a simple macroscopic example. Suppose, as illustrated in Fig. 5.6, that we have hot cup of coffee (body 1) at a temperature of, say, 67° C (so $T_1 = 340.15 \text{ K}$) and a counter-top (body 2) at room temperature of, say, 20° C (so $T_2 = 293.15 \text{ K}$). Imagine that we transfer just 0.01 J of energy from the cup of coffee to the counter-top by briefly laying the cup of coffee on the counter-top. For the purposes of this calculation, we presume that both the cup of coffee and the counter-top are sufficiently large that this small transfer of energy does not appreciably change the temperature of either of them.

So, we have

$$\Delta S_1 = -\frac{0.01}{340.15} \approx -2.94 \times 10^{-5} \text{ JK}^{-1} \quad \text{and} \quad \Delta S_2 = \frac{0.01}{293.15} \approx 3.41 \times 10^{-5} \text{ JK}^{-1}$$

So the total change in entropy by summing these two is

$$\Delta S = (3.41 - 2.94) \times 10^{-5} \approx 4.17 \times 10^{-6} \text{ J K}^{-1}$$

Numerically this verifies that entropy has increased as heat flows from a hotter to a colder body. We can convert to the more fundamental or “statistical” version of entropy by dividing by Boltzmann’s constant, obtaining an increase of entropy of

$$\Delta \sigma = \frac{\Delta S}{k_B} \approx \frac{4.17 \times 10^{-6}}{1.38 \times 10^{-23}} \approx 3.02 \times 10^{17}$$

So, the number of microstates available to the combined system if this heat flows from the hotter to the colder body has increased by a factor

$$\Delta g = \exp(3.02 \times 10^{17}) \approx 10^{(\log_{10} e) \times 3.02 \times 10^{17}} \approx 10^{1.31 \times 10^{17}}$$

which is a truly massive number²⁴. (Note that $10^{10^{17}}$ is 1 followed by 10^{17} zeros.) So, even the transfer of this modest amount of energy from a warm to a cold body corresponds to a massive increase in the number of available microstates.

Suppose for the moment we asked what would happen if this 0.01 J of heat flowed in the other direction, from “cold” at 20° C to “hot” at 67° C, leading to an entropy decrease of this same magnitude. So we would be asking the system to change to a macrostate with a factor of $\exp(3.02 \times 10^{17})$ fewer microstates. Remember that tossing a coin N times leads to 2^N possible outcomes or microstates. For $2^N \approx \exp(3.02 \times 10^{17})$ then

$$N = \log_2(e^{3.02 \times 10^{17}}) = \log_2[(2^{\log_2 e})^{3.02 \times 10^{17}}] = \log_2\{[2^{(1/\log_e 2)}]^{3.02 \times 10^{17}}\} \approx 4.4 \times 10^{17}$$

So, expecting this energy to flow from “cold” to “hot” is like asking for a specific sequence of heads and tails, such as “all heads”, when tossing a coin roughly 4.4×10^{17} times. This will not happen very often. Since the universe is perhaps $\sim 4.4 \times 10^{17}$ seconds old²⁵, this is like tossing a coin once a second since the Big Bang and asking for it to come up heads every time!

To be clear here, we are not saying that we have to successfully randomly manage to generate one specific desired sequence of heads and tails from $\sim 4.4 \times 10^{17}$ possibilities, which would be hard enough. Rather there are $2^{4.4 \times 10^{17}} = 10^{1.31 \times 10^{17}}$ possibilities to choose from!

This calculation illustrates by example the “reason” why heat flows from hot to cold. If there is some mechanism by which energy can flow from one body to another, there are massively more accessible microstates if the energy flows from the hot body to the cold body. To flow in the opposite direction would correspond to the number of accessible microstates decreasing by an equally large factor, the equivalent statistically of

²⁴ We remember, by definition of logarithms to the base a , that $b = a^{\log_a b}$, and for powers that $(a^b)^c = a^{b \times c}$

²⁵ At the time of writing this, the best estimate for the age of the universe (the time since the Big Bang) is ~ 13.8 billion years, or $\sim 4.35 \times 10^{17}$ seconds.

succeeding when randomly trying to choose only one specific state from a very, very large number of possibilities.

The second law of thermodynamics

The second law of thermodynamics is one of the most important principles in physics, and it can be stated in many ways. Many of these statements are related to the behavior of heat engines, which we have not discussed, so we will omit those versions here. The essence of the second law is what we might call the law of increase of entropy, which we could state as

the entropy of a closed system tends to remain constant or to increase.

We have seen above that the entropy of parts of a system can decrease; as heat flows from a hot part of a system to a cold part of a system, the entropy of the hot part does decrease. But, the entropy of the entire system has increased because the entropy of the cold part of the system increased by more than the entropy of the hot part decreased.

It is perhaps best to think of the second law as the rather obvious idea that, given random processes that change the microstate of a system, the system will tend to change towards those macrostates with larger multiplicity – that is, larger numbers of microstates. There are simply many more ways the system will head towards macrostates of larger multiplicity than to macrostates of lower multiplicity. In macroscopic systems, as we have seen by the above calculation on heat transfer, the most-likely macrostates can massively dominate over the others. That is why, for even moderate sizes of (closed) systems, we can state with confidence that the entropy is not going to decrease; such a decrease is just overwhelmingly unlikely²⁶.

The second law is a statistical principle. It is quite possible in small systems for the entropy to decrease sometimes, but those systems need to be really small for there to be much chance of this happening to any significant degree. With the kind of statistical view we are taking here, it is also possible to calculate the possibility of various such random fluctuations in the system away from the “equilibrium” values, and such calculations are an important part of the larger field of statistical mechanics.

Quiz questions

Q5.4.1 We presume we have two systems, system 1 and system 2, each with large fixed numbers N_1 and N_2 , respectively, of quantum particles of the same kind, and

²⁶ You might be worried that somehow the act of you tidying your room, or performing any other such organizing actions that lead to apparently reduced entropy in the world around us, is somehow a violation of the second law of thermodynamics. You can be sure, however, that the increase in the world’s entropy from the energy you expend in the process, which nearly all turns into heat, will vastly exceed any such decrease in entropy in a tidy room! The resulting argument that by not tidying your room, and hence not increasing entropy overall, you are doing the world a favor is, however, not likely to be very convincing on a personal level with your room-mate, and may unfortunately reinforce the entirely inaccurate but widely held belief that scientists and engineers are annoyingly lacking in ordinary social skills.

energies U_1 and U_2 , respectively. The systems are joined by a thermally conducting wall but are otherwise insulated from the environment. Generally, the multiplicities of each system can be written as g_1 and g_2 , respectively. At or near thermal equilibrium, for each of the following, state whether it is necessarily true (“necessarily true” means it has to be true regardless of the specific values of the various parameters, such as N_1 , N_2 , U_1 , and U_2). So, for each of the following, state whether they are True or False (where “True” corresponds to “necessarily true”).

a) $g_1 = g_2$

b) $\left. \frac{1}{g_1} \frac{\partial g_1}{\partial U_1} \right|_{N_1} = \left. \frac{1}{g_2} \frac{\partial g_2}{\partial U_2} \right|_{N_2}$

c) $\left. \frac{\partial g_1}{\partial U_1} \right|_{N_1} = - \left. \frac{\partial g_2}{\partial U_2} \right|_{N_2}$

d) $\left. \frac{\partial \log g_1}{\partial U_1} \right|_{N_1} = \left. \frac{\partial \log g_2}{\partial U_2} \right|_{N_2}$

Q5.4.2 Suppose we have two systems, each initially at different temperatures. When we join two systems using a thermally conducting wall and let them come to thermal equilibrium, we will have minimized the entropy of the combined system. True or False?

Q5.4.3 In a system with a constant number of particles of a given kind, if the rate of change of entropy with energy is increased, that means the temperature has been reduced. True or False?

Q5.4.4 In fundamental units, entropy is a dimensionless quantity. True or False?

Q5.4.5 If we have two separate systems, which we keep separated, then the entropy of the complete system consisting of these two separated systems is the sum of the entropies of the separated systems. True or False?

Q5.4.6 The multiplicity of a system is the logarithm of the entropy of the system. True or False?

Q5.4.7 Consider two systems in a given magnetic field, each with one spin, and starting with system 1 with spin up and system 2 with spin down. We then bring the systems together through a thermally conducting wall.

- How many possible microstates are there for the combined system: 1, 2, or 4?
- How many microstates are there in the most probable macrostate for the combined system?
- What fraction of the possible microstates for the combined system are in the most probable macrostate: 0%, 50% or 100%?

Q5.4.8 A system has a multiplicity of 100. What, approximately, is its entropy in fundamental units: 2, 4.6, -2, or 10? (Note – be careful you take logarithms to

the correct base. Unless otherwise stated, logarithms in fundamental physics and in results in calculus are usually “natural” logarithms, so logarithms to the base e . In this text, we avoid using the “ln” notation for natural logarithms because it is too easily misread, for example as “In”.)

Q5.4.9 1 J of heat energy moves from body 1 at 30°C to another body, body 2 which is at 20°C.

- a) Approximately what is the change in entropy of body 1: $3.3 \times 10^{-3} \left(\frac{J}{K} \right)$, $-3.3 \times 10^{-3} \left(\frac{J}{K} \right)$, or $3.4 \times 10^{-3} \left(\frac{J}{K} \right)$?
- b) Approximately what is the change in entropy of body 2: $3.4 \times 10^{-3} \left(\frac{J}{K} \right)$, $-3.3 \times 10^{-3} \left(\frac{J}{K} \right)$, or $3.3 \times 10^{-3} \left(\frac{J}{K} \right)$?
- c) Approximately what is the change in entropy overall: $6.8 \times 10^{-3} \left(\frac{J}{K} \right)$, $6.6 \times 10^{-3} \left(\frac{J}{K} \right)$, $-0.1 \times 10^{-3} \left(\frac{J}{K} \right)$ or $0.1 \times 10^{-3} \left(\frac{J}{K} \right)$?

Q5.4.10 As a result of some heat transfer, the entropy of a body increases by 100 in fundamental units. By approximately which factor has the multiplicity of the body increased: $10^{43.43}$, $10^{3 \times 10^{23}}$, or 1000?

Q5.4.11 As a result of some heat transfer, the entropy of a body increases by 1 J/K. By approximately which factor has the multiplicity of the body increased: $10^{3.2 \times 10^{22}}$ or 10^{43} ?

Problems

5.4.1 Suppose we have a set of 100 spins, 55 of which are spin-up, and 45 of which are spin-down, and that is all we know about this particular “macrostate” of the system.

- (i) What is the multiplicity of this macrostate (at least approximately)?
- (ii) What is the corresponding entropy of this macrostate?
 - (a) in “fundamental” units?
 - (b) in $J K^{-1}$?

5.4.2 In thermodynamics, if an amount of heat energy U passes into a body at temperature T , then the change of entropy of the body is $\Delta S = U / T$. For $U = 5 J$ and a temperature of 20° C,

- (i) what is the change of entropy of the body, in thermodynamic units?
- (ii) by what factor has the multiplicity of the accessible microstates changed?

5.4.3* Consider 1 kg of water, which in this question may be in solid (ice), liquid (liquid water), or gas (water vapor) form. Find the change in entropy (in thermodynamic units) of the water in each of the following cases.

- (a) 1 kg of ice melts at a temperature of $T = 0^\circ\text{C}$. (Take the latent heat of fusion (the heat energy involved in melting or freezing a kilogram) to have a magnitude $3.34 \times 10^5 \text{ J kg}^{-1}$ under the conditions for this experiment. Note that heat has to be added to ice to melt it.)
- (b) 1 kg of water vapor at $T = 100^\circ\text{C}$ condenses into liquid water. (Take the latent heat of vaporization (the energy involved in boiling or condensing a kilogram) to have a magnitude $2.26 \times 10^6 \text{ J kg}^{-1}$ under the conditions for this experiment. Note that heat has to be added to liquid water to boil (vaporize) it.)

5.4.4 An ideal gas is defined as one in which all collisions between atoms or molecules are perfectly elastic and in which there are no intermolecular attractive forces. It is like the kind of “gas” that would correspond to a large collection of very small “billiard balls” bouncing elastically off one another inside a box. Such an ideal gas is a useful first model for considering the behavior of gasses. The kinetic energy of such an ideal gas at a given temperature T is $U = (3/2)Nk_B T$ where N is the (fixed) number of gas particles. For the purposes of this question, we can consider this U to be the “heat” energy of the gas.

We want to understand, essentially, what is the increase in entropy ΔS of such a gas if, starting at some energy U_i in the gas, we slowly add some total amount ΔU of heat to it. Note that, as we slowly add this heat, we are also slowly going to change the temperature of the gas, which complicates the solution somewhat. Now, we know the (thermodynamic) relation between small changes δS in entropy and changes δU in (heat) energy at a given temperature T , for example,

as the derivative $\frac{1}{T} = \left(\frac{\partial S}{\partial U} \right) \Big|_N$ in Eq. (5.42). So, we are going to have to integrate

this derivative to get the total corresponding change in entropy ΔS as we add this heat energy ΔU .

Now, because of the very simple relation $U = (3/2)Nk_B T$ in this ideal gas, or

equivalently, $T = \frac{2}{3Nk_B} U$, we can rewrite this derivative in this case as

$$\left(\frac{\partial S}{\partial U} \right) \Big|_N = \frac{3}{2} \frac{Nk_B}{U}.$$

So now, starting from some energy U_i and ending at some energy $U_i + \Delta U$, we could now deduce an expression for the corresponding change in entropy by integrating this derivative.

For this question, the answer we want is not to state this resulting entropy change itself, but instead to deduce by what factor $g(\Delta U)$ the multiplicity of the gas

changes as a function of this amount of (heat) energy ΔU added to the gas, with a given number of particles N .

5.5 Carnot efficiency limit for heat engines 18c

A heat engine is a device that takes heat energy, say U_H , from a hot source or “reservoir”, performs some useful work of energy W (so, turning a shaft to raise a weight, for example, or generating electricity), and dumps some heat energy, say U_C , into a cold sink or reservoir. An idealized heat engine is shown in Fig. 5.7.

Such heat engines are very common. All internal combustion piston engines (gasoline or diesel) are heat engines. The hot reservoir is the hot gas in the cylinders of the engine after the fuel has combusted. The cold reservoir is usually a “radiator”, often based on water heated in the cylinder block of the engine and then circulated to the radiator where the heat is dissipated to the atmosphere. A steam (piston) engine is also obviously a heat engine; in this case the combustion is external, in some fire that heats a boiler, for example, with the hot water or steam being the hot “reservoir”.

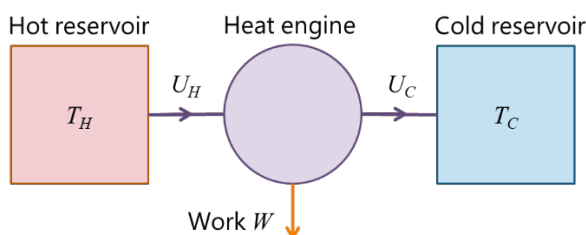


Fig. 5.7. An idealized heat engine that takes heat energy U_H in from a hot reservoir at temperature T_H , performs work of energy W , and dumps heat energy U_C into a cold reservoir at temperature T_C .

The Stirling engine is an entirely sealed piston heat engine, requiring only external hot and cold reservoirs²⁷. Jet engines are gas turbine (heat) engines. Most electricity generation from fuels like coal, gas or nuclear energy is by steam turbines, which are also heat engines. A turbine that generates electrical power is obviously a heat engine since the electrical power is generated based on the mechanical work that the engine creates to power a dynamo or some kind. Generally, any engine that generates electrical power from heat is also a heat engine, however, because electrical power can be converted to mechanical work with 100% efficiency in principle. So, solar cells are also heat engines, with the hot reservoir being the sun, even though they do not directly generate mechanical motion.

With two simple principles – conservation of energy (which is also essentially the First Law of Thermodynamics), and the requirement the entropy cannot decrease overall

²⁷ Stirling engines can work with only small temperature differences. Toy Stirling engines can run when you put them on top of your hot cup of coffee or tea, and are fun to play with.

(which is essentially the Second Law of Thermodynamics), we can calculate a limit to the energy efficiency of any such heat engine; this limit is known as the Carnot limit²⁸.

We want to calculate a limit to how efficient such a heat engine could possibly be, so we presume no energy is wasted in any other processes, such as friction or heat loss to any other heat sink. So, by conservation of energy,

$$U_C = U_H - W \quad (5.52)$$

The entropy change of the hot reservoir is

$$\Delta S_H = -\frac{U_H}{T_H} \quad (5.53)$$

This is a decrease in entropy of the hot reservoir because heat energy is flowing out of the hot reservoir, hence the minus sign. The entropy change of the cold reservoir is

$$\Delta S_C = \frac{U_C}{T_C} = \frac{U_H - W}{T_C} \quad (5.54)$$

This is an increase in entropy of the cold reservoir because heat energy is flowing into the cold reservoir. By presumption, entropy overall cannot decrease, so

$$\Delta S_H + \Delta S_C \geq 0 \quad (5.55)$$

So

$$-\frac{U_H}{T_H} + \frac{U_H - W}{T_C} \geq 0 \quad (5.56)$$

Rearranging gives

²⁸ Carnot devised this limit in 1824. At that time, the concept of entropy was not yet developed, so Carnot used a different and ingenious argument. He proposed a hypothetical engine operating in a cycle based on expanding gas in a cylinder pushing a piston, with appropriate “isothermal” (i.e., same temperature) and “adiabatic” (i.e., with no heat flow) steps. The approach involves the cylinder connected to a hot reservoir, then detached from all reservoirs, and then connected to a cold reservoir, with similar steps then repeated in reverse to get us back to where we started, completing a cycle in which some work has been done by the cylinder’s piston, with heat flow out of the hot reservoir and heat flow into the cold reservoir. Such an approach can be made entirely reversible, allowing either a heat engine that converts heat flow from hot to cold, with some efficiency, into work, or by putting in work, to refrigerate the cold reservoir and put heat into the hot reservoir. We will not give the details of this argument here, though it is an easy one to find in the literature. Carnot’s approach starts, then, by constructing such a hypothetical and completely reversible cycle. In modern terms, such reversibility is associated with having no change in entropy. In Carnot’s argument, to conclude that the efficiency of this cycle is the best possible in a heat engine, one only has to believe that no machine or process could allow the flow of heat from cold to hot, with no other net change in the system. If we had a heat engine more efficient than a Carnot one, then we could use it to drive a Carnot engine in reverse (so running the Carnot engine as a refrigerator), and the net result would be some net transfer of heat from cold to hot by the time we had completed the full cycle of both machines. Hence by *reductio ad absurdum*, no such more efficient engine can be constructed.

$$U_H \left(\frac{1}{T_C} - \frac{1}{T_H} \right) \geq \frac{W}{T_C} \quad (5.57)$$

Defining the engine's efficiency η_{engine} as the work energy out per unit heat energy in (presuming we have to keep replenishing the hot reservoir to make up for the heat extracted from it)

$$\eta_{engine} \equiv \frac{W}{U_H} \leq 1 - \frac{T_C}{T_H} \quad (5.58)$$

which is the Carnot efficiency limit for the heat engine. For given temperatures, nothing we can do can make a heat engine more efficient than this. To do so would require violating either the First or Second Laws of Thermodynamics.

This is a very serious and real limit. We can never convert heat energy continuously to work at 100% efficiency. To get even reasonably efficient engines, we need a very hot "hot" reservoir and/or a very cold "cold" reservoir.

Quiz questions

- Q5.5.1 When we run an ideal heat engine to perform work, the entropy of both the hot and cold reservoirs increases. True or False?
- Q5.5.2 The Carnot efficiency of an ideal heat engine will increase if we increase the temperature of the hot reservoir. True or False?
- Q5.5.3 The Carnot efficiency of an ideal heat engine will increase if we increase the temperature of the cold reservoir. True or False?

Problems

- 5.5.1 What is the maximum possible efficiency of a solar cell at converting the sun's energy to electricity, presuming the sun has a temperature of 5800 K, and the earth is the "cold" reservoir at a temperature of 300 K?
- 5.5.2 Perform a similar analysis to that for the heat engine, but for a refrigerator that is to extract heat energy from the cold reservoir at temperature T_C by performing work and delivering all the energy to the hot reservoir at temperature T_H .
- Defining the efficiency $\eta_{refrigerator}$ as the ratio of the magnitude U_C of the heat energy extracted from the cold reservoir divided by the magnitude W of the work energy put in to the "refrigerator" engine, derive an expression for the maximum efficiency of the refrigerator in terms of T_C and T_H .
 - What is the maximum efficiency if the cold reservoir is to be at 0° C (so, 273.15 K) and the hot reservoir is to be at 30° C (so 303.15 K)?
 - If the refrigerator is sitting in the middle of your room, can you continuously cool the room by opening the refrigerator door? (Justify your answer.)

- (iv) What is the maximum efficiency if the cold reservoir is to be at 4.2 K (the boiling temperature of liquid helium, and a common cryogenic temperature) and the hot reservoir is to be at 30° C (so 303.15 K)?
 - (v) You notice that the best logic device you have available at room temperature takes 40 aJ to switch states. You want to save energy in information processing so you look for devices that take less energy. You notice that, if you cool this device down to 4.2 K, it can switch with only 4 aJ of energy. You pitch investors to back your proposal to develop this cooled technology to save energy in information processing. Presuming that the investors have taken thermal physics courses (and therefore must be both very intelligent and highly educated), should they invest in your proposal? Justify your answer (scientifically!).
- 5.5.3* Perform a similar analysis to that for the heat engine, but for a “heat pump”. A heat pump is a device in which we put in work W to heat our house by adding heat energy U_H to the inside of our house at temperature T_H (which is then the “hot” reservoir) while simultaneously cooling (or extracting heat from) a cold reservoir T_C (which is the environment outside our house). Its efficiency can be defined as the ratio η_{heatpump} of the heat energy U_H delivered to the house at temperature T_H to the work energy W put into the heat pump
- (i) Derive an expression for the maximum efficiency of the heat pump in terms of T_C and T_H .
 - (ii) What is the maximum efficiency of the heat pump, defined as $\eta_{\text{heatpump}} = U_H / W$, if the outside temperature is 0° C (273.15 K) and the inside of the house is at 25° C (298.15 K)

5.6 The Boltzmann factor

We are often interested in knowing about the properties of a system that is being held at some specific temperature, such as how particles are distributed among the various possible available states. That is a common situation in real life, for example for thermal distributions of electrons in semiconductor electronic and optoelectronic devices.

One good way of describing such a situation is to imagine that we have some system of moderate size, which we will call system M , and which is being held in thermal contact with a very large system, which we call the reservoir, R . Energy can be exchanged through the walls between the system M and the reservoir R , but we presume we are only interested for the moment in situations where no particles can be exchanged between the system M and the reservoir R .

The energy in the total system, which consists of our system M and the reservoir R , is presumed to be some fixed amount U_o . We presume that our system M has an amount of energy ε , so the energy of the reservoir R is $U_o - \varepsilon$, as illustrated in Fig. 5.8

Our goal here is to establish the probability that the system M is in some particular quantum state (which therefore will be a microstate of M) given that we have a particular temperature $\tau (\equiv k_B T)$ for the total system. Consider, then, two possible

quantum states of the system M , state 1 and state 2, with associated energies ε_1 and ε_2 . The probability P_1 that the system M is in state 1 is proportional to the multiplicity of the reservoir R when it has energy $U_o - \varepsilon_1$; this multiplicity is simply the number of ways the total system can exist in which system M is in state 1.

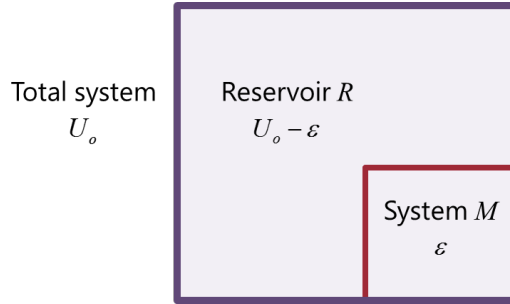


Fig. 5.8. A total system with energy U_o , divided into a large reservoir R , with energy $U_o - \varepsilon$, and a system M with energy ε . The system M can exchange energy with the reservoir through its thermally conducting walls.

So, with a similar argument for state 2 and the corresponding probability P_2 of the system M being in state 2, we have

$$\frac{P_1}{P_2} = \frac{\text{Multiplicity of } R \text{ at energy } U_o - \varepsilon_1}{\text{Multiplicity of } R \text{ at energy } U_o - \varepsilon_2} \quad (5.59)$$

Note here that we are not discussing the probabilities that the system M has energy ε_1 or ε_2 ; there could possibly be many different quantum states (and hence microstates) that have the same energy ε_1 or ε_2 ; we are discussing the probabilities that the system is in a specific quantum state with energy ε_1 or in another specific quantum state with energy ε_2 .

Now, with $\sigma_R(U)$ as the entropy of the reservoir at some energy U , we know that

$$\text{Multiplicity of } R \text{ at energy } U_o - \varepsilon_1 = \exp[\sigma_R(U_o - \varepsilon_1)] \quad (5.60)$$

and similarly

$$\text{Multiplicity of } R \text{ at energy } U_o - \varepsilon_2 = \exp[\sigma_R(U_o - \varepsilon_2)] \quad (5.61)$$

So, from Eq. (5.59), we have

$$\frac{P_1}{P_2} = \frac{\exp[\sigma_R(U_o - \varepsilon_1)]}{\exp[\sigma_R(U_o - \varepsilon_2)]} = \exp[\sigma_R(U_o - \varepsilon_1) - \sigma_R(U_o - \varepsilon_2)] \quad (5.62)$$

Now, we are presuming that, compared to the energy of the very large reservoir, the energy ε of the specific state of the system M is very small. So, we can usefully expand the entropy of the reservoir in a Taylor series about the point U_o . Hence we obtain, formally,

$$\sigma_R(U_o - \varepsilon) = \sigma_R(U_o) - \varepsilon \left(\frac{\partial \sigma_R}{\partial U} \right) \Big|_{U=U_o} + \frac{1}{2} \varepsilon^2 \left(\frac{\partial^2 \sigma_R}{\partial U^2} \right) \Big|_{U=U_o} + \dots \quad (5.63)$$

Note, implicitly, all these derivatives are being taken at constant numbers of particles in both the reservoir and the system M , and we are presuming other possible variables, like the volumes of the systems, are being held constant also. Given these variables being held constant, we know from Eq. (5.40) that here we have

$$\left(\frac{\partial \sigma_R}{\partial U} \right) \Big|_{U=U_o} = \frac{1}{\tau} \quad (5.64)$$

where $\tau (\equiv k_B T)$ is the temperature of the reservoir.

The reservoir is taken to be very large by definition. Changing the reservoir's energy by any small amount therefore makes negligible difference to its temperature. If we do find that changing the reservoir's energy by some amount ε of interest to us does change the temperature measurably, then we simply make the reservoir bigger until that temperature change is negligible. So, we presume for no energy change of interest to us does the temperature of the reservoir change appreciably, and hence the higher derivative terms in the expansion Eq. (5.63) can all be made to vanish by choosing the reservoir sufficiently large²⁹. Hence, retaining only the first-order term in ε , we can write

$$\sigma_R(U_o - \varepsilon) = \sigma_R(U_o) - \frac{\varepsilon}{\tau} \quad (5.65)$$

Using this expression in Eq. (5.62), we therefore have

$$\frac{P_1}{P_2} = \frac{\exp(-\varepsilon_1 / \tau)}{\exp(-\varepsilon_2 / \tau)} = \exp \left[-\frac{(\varepsilon_1 - \varepsilon_2)}{\tau} \right] \equiv \exp \left[-\frac{(\varepsilon_1 - \varepsilon_2)}{k_B T} \right] \quad (5.66)$$

So, for a system M in thermal contact with a large reservoir at temperature $\tau (\equiv k_B T)$, the relative probability, under conditions of constant numbers of particles, of the system M being in a specific quantum state of energy ε_1 rather than one of energy ε_2 is given by Eq. (5.66). This factor that expresses this relative probability of occupation of states separated by energy ε in thermal equilibrium, is called

$$\text{the Boltzmann factor, } \exp \left(-\frac{\varepsilon}{k_B T} \right)$$

where we have written it in conventional units.

The Boltzmann factor and others closely related to it appear in a wide range of situations, and form the basis for the distributions of particles among quantum mechanical states in systems in thermal equilibrium.

²⁹ Note that $\partial^2 \sigma_R / \partial U^2$ is the rate of change of $1/(\text{the reservoir's temperature})$ with energy, which we can make arbitrarily small by choosing the reservoir sufficiently large.

Quiz questions

Q5.6.1 An atom is inside a box that is connected through a thermally conducting wall to a large reservoir that is at room temperature. Presume we can take room temperature to correspond to $k_B T = 25 \text{ meV}$. Suppose the atom has three energy levels. Level 1 has an energy that we choose arbitrarily to be zero. Level 2 has an energy of 25 meV. Level 3 has an energy of 75 meV. The atom may also have many other possible energy levels.

- What is the ratio $\frac{P_2}{P_1}$ of the probability P_2 of the atom being in level 2 compared to the probability P_1 it is in level 1: $1/e$, e , $1/3$, or $\exp(0.025)$?
- What is the ratio $\frac{P_3}{P_2}$ of the atom being in level 3 compared to the probability P_2 that it is level 2: $1/e$, $\exp(0.050)$, e^3 , $1/e^2$?
- What is the ratio $\frac{g_1}{g_2}$ of the multiplicity of the reservoir when the atom is in state 1 (with reservoir multiplicity g_1) compared to the when the atom is in state 2 (with reservoir multiplicity g_2): $1/e$, e , $\log e$ or 0?

Problems

5.6.1* Consider a hydrogen atom that is in thermal equilibrium with a large heat reservoir. What is the relative probability of finding the electron in a specific $n = 6$ state compared to finding it in a specific $n = 5$ state at each of the following temperatures?

- Room temperature $T = 293 \text{ K}$
- Sun temperature $T = 5780 \text{ K}$

5.6.2 Consider a large number of hydrogen atoms in thermal equilibrium at a temperature of 5800 K (approximately, the temperature of the sun). N_1 atoms will be in a 1s state, and N_2 atoms will be in a 2s state. What value would we expect for the ratio N_2 / N_1 ? [For simplicity, you can consider atoms of just one specific electron spin and one specific proton spin, though including all possible spin states actually makes no difference to this problem.]

5.7 Chemical potential and the Gibbs factor

To be most useful, we should extend this discussion by one more step, which is to include the consequences of allowing particles to move between the system M and the reservoir R .

Suppose we had two systems that were coupled through a permeable membrane instead of a thermally conducting wall. So now we could have particles move from one side to the other through the membrane. Such “diffusion” processes can occur even if the temperature is the same on both sides. We can argue now that the joint system should now settle down, in what we can call diffusive equilibrium, to some situation that maximizes entropy overall.

Chemical potential

Because we are interested in the change in entropy that arises just from the transfer of particles, we will presume for the moment that this movement of particles does not itself change the energy of either system. In our example spin system, this would be the case if we turned off the magnetic field, for example, since then every spin would have zero energy anyway. So we can examine a relation like the one we put together for the change in entropy with changes in energies at constant particle number, Eq. (5.47), but now thinking of the change of entropy with changes of particle numbers N_1 and N_2 at constant energies U_1 and U_2 for systems 1 and 2, that is

$$\Delta\sigma = \left(\frac{\partial\sigma_1}{\partial N_1} \right)_{U_1} \Delta N_1 + \left(\frac{\partial\sigma_2}{\partial N_2} \right)_{U_2} \Delta N_2 \quad (5.67)$$

Here, we presume conservation of particle number overall, so if a small number ΔN of particles moves from system 1 to system 2, then $\Delta N_1 = -\Delta N$ and $\Delta N_2 = \Delta N$. Presuming we are at diffusive equilibrium, which we take to be the situation of maximum entropy, then any such small transfer should result in vanishing change in entropy, so, at diffusive equilibrium we should have

$$0 = - \left(\frac{\partial\sigma_1}{\partial N_1} \right)_{U_1} \Delta N + \left(\frac{\partial\sigma_2}{\partial N_2} \right)_{U_2} \Delta N \quad (5.68)$$

which means that, at diffusive equilibrium

$$\left(\frac{\partial\sigma_1}{\partial N_1} \right)_{U_1} = \left(\frac{\partial\sigma_2}{\partial N_2} \right)_{U_2} \quad (5.69)$$

So the quantity represented by such a derivative for a system is the quantity that is equalized under diffusion of particles from one system to the other.

Conventionally in thermodynamics, we use the quantity called the chemical potential, which is defined as

$$\mu_C = -\tau \left(\frac{\partial\sigma}{\partial N} \right)_{U} \quad (5.70)$$

which is simply (minus) the (fundamental) temperature times this derivative³⁰. Given that we are operating at some temperature, then we can say that diffusive equilibrium for some kind or “species” of particle between two systems or bodies corresponds to both systems or bodies having the same chemical potential for that species. Here, of course, so far we have only been considering one kind or species of particle. We can generalize this discussion to multiple different kinds of particles, in which case each kind of particle can have its own chemical potential.

An equivalent definition of the chemical potential

The way we write the chemical potential above, as in Eq. (5.70), is the one we use in this chapter as we discuss thermal distributions. There is another, equivalent definition that we need later when discussing electronic and optoelectronic devices; in particular, it allows us to understand what happens to the chemical potential when we apply voltages to devices. For completeness, we formally derive this alternative definition here, though we will not use it until Chapter 6³¹. This version can be written as

$$\mu_C = \left(\frac{\partial U}{\partial N} \right) \bigg|_S \quad (5.71)$$

That is, the chemical potential is the rate of change of the energy of the system per added particle of the given species, when we hold the entropy constant during this addition. This is the definition of chemical potential as originally introduced by Gibbs. He proposed the chemical potential to be the increase in energy per unit quantity of substance added, under conditions of constant entropy. This concept of an amount of energy as a function of the amount of some chemical substance justifies the idea of calling it a “chemical potential”.

It is, however, by no means obvious physically why Gibbs’ definition, Eq. (5.71) would be equivalent to the chemical potential definition in Eq. (5.70). We can now prove the equivalence of Eqs. (5.70) and (5.71) algebraically; like many relations involving partial derivatives in thermodynamics, however, it can be quite difficult to understand the equivalence of such expressions in a direct physical way.

To see the relation between the two definitions, consider the differential of entropy σ for a system with energy U and particle number N . Then, choosing to consider a situation of constant entropy, we would have

$$d\sigma = \left(\frac{\partial \sigma}{\partial U} \right) \bigg|_N dU + \left(\frac{\partial \sigma}{\partial N} \right) \bigg|_U dN = 0 \quad (5.72)$$

³⁰ Typically, chemical potential is just written as μ , without the subscript “C” we have used here, but we introduced the subscript to avoid confusion with other uses of μ in this text.

³¹ Video lecture discussion of this section is also postponed till Chapter 6.

Now, we are interested in finding a partial derivative at constant entropy, as in Eq. (5.71). So indeed the fact that we set $d\sigma = 0$ in Eq. (5.72) means that, if we essentially³² divide both sides by dN , the resulting ratio of dU over dN is one taken at constant entropy (which we can view as either σ or S , since they are just related by a constant).

So, technically that ratio is the partial derivative $\left. \frac{\partial U}{\partial N} \right|_S$. Hence we have, after rearrangement

$$\left(\frac{\partial \sigma}{\partial U} \right) \Big|_N \left(\frac{\partial U}{\partial N} \right) \Big|_S = - \left(\frac{\partial \sigma}{\partial N} \right) \Big|_U \quad (5.73)$$

So, using our definition of (fundamental) temperature τ , Eq. (5.40) (which is $1/\tau = (\partial \sigma / \partial U) \Big|_N$), we prove³³ the equivalence

$$\left(\frac{\partial U}{\partial N} \right) \Big|_S = -\tau \left(\frac{\partial \sigma}{\partial N} \right) \Big|_U \quad (5.74)$$

As we mentioned, this relation will be useful to us later when thinking about semiconductor devices. If we take some piece of material containing electrons and

³² The idea of “dividing by dN ” to form a derivative like this is technically what is sometimes called an “abuse of notation”. This Leibniz notation of writing derivatives as ratios of infinitesimal quantities (as in dy/dx) tempts us to do this, and we can often get away with it,

for example in a “chain rule” where we write $\frac{dy}{dz} \frac{dz}{dx} = \frac{dy}{dx}$, and just “cancel” the dz on the top

and bottom lines. (Here z had to be a function of just x , i.e., $z(x)$.) For expressions such as Eq. (5.72), we can be more formally correct in our proof if we rewrite first of all in terms of actual

small changes $\Delta \sigma$, ΔU , and ΔN , which gives $\Delta \sigma = \left(\frac{\partial \sigma}{\partial U} \right) \Big|_N \Delta U + \left(\frac{\partial \sigma}{\partial N} \right) \Big|_U \Delta N = 0$. Then

we can quite legally divide by ΔN , giving $\left(\frac{\partial \sigma}{\partial U} \right) \Big|_N \frac{\Delta U}{\Delta N} \Big|_S + \left(\frac{\partial \sigma}{\partial N} \right) \Big|_U = 0$. We then formally

take the limit as ΔN and the resulting ΔU tend to zero, giving $\left(\frac{\partial \sigma}{\partial U} \right) \Big|_N \left(\frac{\partial U}{\partial N} \right) \Big|_S + \left(\frac{\partial \sigma}{\partial N} \right) \Big|_U = 0$

or, equivalently, (5.73).

³³ There are various different equivalent ways of deriving and of defining chemical potential. Some of these discuss derivatives of other quantities, sometimes known as thermodynamic potentials, such as Helmholtz free energy, Gibbs free energy, and enthalpy. Though those thermodynamic potentials are very convenient for particular important classes of problems, they are not as fundamental as energy and entropy; thermodynamics can in principle be understood without them. We restricted ourselves here to using just energy and entropy, and we performed our derivation of chemical potential without using those other quantities. As a result, our derivation above may look different from many others – for example, it is common to use the Helmholtz free energy in the proof of Eq. (5.74) – and our (somewhat original) derivation here is arguably more compact than most.

change the voltage applied to it by some amount V , we correspondingly are just adding the same amount of energy (of magnitude $e \times V$, where e is the magnitude of the electron charge) to every electron without making any other changes. So any derivative like $\left. \frac{\partial U}{\partial N} \right|_s$ just increases by that same additive amount. Hence changing voltage just changes the chemical potential by an amount of magnitude $e \times V$. The chemical potential for electrons in semiconductors is known more commonly as the “Fermi level”, so this tells us that the Fermi level just moves with the applied voltage. We use that idea very frequently in discussing electronic and optoelectronic devices, and we will return to this point later.

The Gibbs factor

Now that we are considering the exchange of both particles and energy, we can derive another useful factor, called the Gibbs factor, that is analogous to the Boltzmann factor that arose when we considered only the exchange of energy.

Consider now a system M in thermal and diffusive contact with a large reservoir R . Now, in addition to being able to transfer energy between the system and the reservoir, we can also transfer particles. The entire closed system of reservoir R and system M has N_o identical particles, and has energy U_o . When the system M has N particles, then the reservoir has $N_o - N$ particles, and when the system M has energy ε_M , the reservoir has energy $U_o - \varepsilon_M$.

Just as for the derivation of the Boltzmann factor, we consider the system M to be in a particular state, in which it has energy ε_M and N particles. We consider two specific microstates.

State 1 is a microstate in which system M has energy ε_1 and N_1 particles

State 2 is a microstate in which system M has energy ε_2 and N_2 particles

Note now when we are considering these two different microstates, we are not just thinking about two different quantum mechanical states of a given particle such as an atom or an electron. We are also allowing the particles to move in and out of the small system M . So the description of a specific “state” (actually a microstate) of system M includes not just the quantum mechanical states occupied by the particles inside system M but also whether or not the particles are even in system M . So one microstate of system M might have 3 hydrogen atoms, with 2 in a specific 1s state, and one in a specific 2p state; another microstate, with a different energy, might have 5 hydrogen atoms, all in a specific 1s state. With the derivation of the Boltzmann factor, we did not necessarily notice this difference between a microstate of the system M and the quantum mechanical state of a particle because we were not allowing any exchange of particles between the system M and the reservoir, but in considering the Gibbs factor here, we need this broader understanding of a state (technically, a microstate) of the system M .

The ratio of the probabilities of M being in state 1 or state 2 is

$$\frac{P(N_1, \varepsilon_1)}{P(N_2, \varepsilon_2)} = \frac{g(N_o - N_1, U_o - \varepsilon_1)}{g(N_o - N_2, U_o - \varepsilon_2)} \quad (5.75)$$

where g in each case is the corresponding multiplicity of the whole system when system M is in the given state³⁴. We know, by definition, that the multiplicity g is just the exponential of the corresponding entropy σ , so

$$\frac{P(N_1, \varepsilon_1)}{P(N_2, \varepsilon_2)} = \frac{\exp[\sigma(N_o - N_1, U_o - \varepsilon_1)]}{\exp[\sigma(N_o - N_2, U_o - \varepsilon_2)]} = \exp(\Delta\sigma) \quad (5.76)$$

where

$$\Delta\sigma = \sigma(N_o - N_1, U_o - \varepsilon_1) - \sigma(N_o - N_2, U_o - \varepsilon_2) \quad (5.77)$$

Just as we did for the Boltzmann factor derivation, we expand $\sigma(N_o - N, U_o - \varepsilon)$ in a Taylor series, now about the values N_o and U_o , with the difference that this time we need to expand in two variables, N and ε , to obtain

$$\sigma(N_o - N, U_o - \varepsilon) = \sigma(N_o, U_o) - N \left(\frac{\partial \sigma}{\partial N_o} \right)_{U_o} - \varepsilon \left(\frac{\partial \sigma}{\partial U_o} \right)_{N_o} + \dots \quad (5.78)$$

Assuming that the reservoir is very large, we may neglect all higher order terms in the expansion, just as we did in the Boltzmann factor derivation, to obtain, in the case of a very large reservoir

$$\Delta\sigma = -(N_1 - N_2) \left(\frac{\partial \sigma}{\partial N_o} \right)_{U_o} - (\varepsilon_1 - \varepsilon_2) \left(\frac{\partial \sigma}{\partial U_o} \right)_{N_o} \quad (5.79)$$

So, using our definitions of temperature, Eq. (5.40), and chemical potential Eq. (5.70), we have

$$\Delta\sigma = \frac{(N_1 - N_2)\mu_C}{\tau} - \frac{(\varepsilon_1 - \varepsilon_2)}{\tau} \quad (5.80)$$

Finally, from Eq. (5.76)

$$\frac{P(N_1, \varepsilon_1)}{P(N_2, \varepsilon_2)} = \frac{\exp[(N_1\mu_C - \varepsilon_1)/\tau]}{\exp[(N_2\mu_C - \varepsilon_2)/\tau]} \quad (5.81)$$

The factor that expresses this relative probability of occupation of states separated by energy ε and by population difference N is called

$$\text{the Gibbs factor, } \exp\left[\frac{(N\mu_C - \varepsilon)}{\tau}\right]$$

³⁴ Each g here might look as if it is only the multiplicity of the reservoir, but since the system M is in a specific microstate, and hence has multiplicity 1, these are the multiplicities of the combined system.

which is the un-normalized probability that system M is in a state of energy ε and particle number N .

Quiz questions

- Q5.7.1 Suppose we have some large number of particles, all of the same energy, in two boxes, 1 and 2, and that we have allowed the particles to move between the boxes so that we have reached diffusive equilibrium. If we now move a particle from box 1 to box 2, state whether each of the following is True or False.
- The entropy of both boxes will decrease
 - The entropy of both boxes will increase
 - The entropy of one box will increase and that of the other will decrease
 - The overall entropy of the system of two boxes will remain at least approximately the same
- Q5.7.2 Suppose we have some atoms that can be in state 1, with energy $E_1 = 0$ or state 2, with energy $E_2 = 25 \text{ meV}$ (and we take 25 meV to correspond to $k_B T$ at room temperature). We have a small system that may contain such atoms, and the atoms are free to move through a permeable wall to a large thermal reservoir that also contains a very large number of these atoms. The permeable wall is also thermally conductive, so energy can be exchanged with the reservoir. We presume the small system and the reservoir are in both thermal and diffusive equilibrium. Suppose first that the density of particles on both sides is such that the chemical potential $\mu_c = 25 \text{ meV}$. Consider two probabilities: The probability P_A that there is one atom in the small system and it is in state 1; and the probability P_B that there two atoms in the small system and they are both in state 2. What is the ratio P_B / P_A ? Choose one of the following.
- e
 - 1
 - $1/e$
 - 0

5.8 Distribution functions

The Fermi-Dirac distribution

Suppose we are interested in fermions, which we know obey the Pauli exclusion principle; in practice here, the fermions of most interest to us will be electrons. So, we know that a particular quantum-mechanical state, such as a state of a particular $\mathbf{k}$ in a band in a semiconductor, or an orbital in an atom, can only be occupied by one fermion of a given spin. We will consider the particles of a given spin, and will derive a form for their distribution with temperature, based on the Gibbs factor. So, suppose now that our system M is in just one particular possible such state for a fermion of a given spin.

So we only have two possibilities for N for this state. Either $N = 1$, with an energy we will call ε , or $N = 0$, with an energy³⁵ 0.

Now, the Gibbs factor just represents relative probabilities of two different possibilities. To get to absolute probability, we just need to divide by the sum of the relative probabilities for the two different possibilities. That will ensure that all the absolute probabilities add up to one, as required³⁶.

Such a sum of relative probabilities over all the possible states is known as a “partition function”, often denoted by Z . For a single state for a fermion, the partition function, is particularly simple, just having two possible terms in it.

The Gibbs factor for energy 0 and $N = 0$ is $\exp[(0 \times \mu_C - 0) / \tau] = 1$, and that for energy ε and $N = 1$ is $\exp[(1 \times \mu_C - \varepsilon) / \tau]$, so the sum of these relative probabilities, the partition function Z , is

$$Z = 1 + \exp\left[\frac{(\mu_C - \varepsilon)}{\tau}\right] \quad (5.82)$$

To get the absolute probability for occupation of this state of energy ε , then, we divide the relative probability (the Gibbs factor) for the occupation of this state of energy ε – that is, $\exp[(\mu_C - \varepsilon) / \tau]$ – by the sum of these two relative probabilities – that is, the partition function Z above. So, the (absolute) probability that this state is occupied by a fermion (such as an electron) is

$$P(N = 1, \varepsilon) \equiv f(\varepsilon) = \frac{\exp[(\mu_C - \varepsilon) / \tau]}{1 + \exp[(\mu_C - \varepsilon) / \tau]} \quad (5.83)$$

or, multiplying top and bottom by $\exp[(\varepsilon - \mu_C) / \tau]$, since

$$\exp(x) \times \exp(-x) = \exp(x - x) = \exp(0) = 1 \quad (5.84)$$

equivalently we can write this as

$$f_{FD}(\varepsilon) = \frac{1}{\exp[(\varepsilon - \mu_C) / \tau] + 1} \quad (5.85)$$

³⁵ The absolute energies of these states do not matter since we will only be considering energy differences, so the choice of this energy to be zero is of no particular significance as long as the energy difference here is ε .

³⁶ Suppose, for example, that when I buy shirts, for each 4 white shirts I buy, I always buy 3 blue shirts and one red shirt. I have bought so many shirts, however, that I can't remember how many I have. I want to know the probability that, when I select a shirt at random, it is a blue shirt. The “unnormalized” or “relative” probabilities here are obviously: white – 4; blue – 3; red – 1. Obviously also, these relative probabilities do not add up to one (in fact, in this case, they are each 1 or larger). I can find the actual (“normalized”) probability by dividing by the sum of all the “unnormalized” or “relative” probabilities. So, the probability of choosing a blue shirt at random is $3 / (4 + 3 + 1) = 3 / 8$.

which is called the Fermi-Dirac distribution, as plotted in Fig. 5.9 and Fig. 5.10.

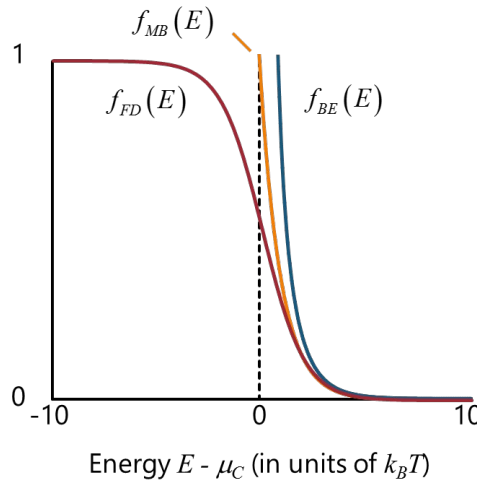


Fig. 5.9. Fermi-Dirac (FD), Maxwell-Boltzmann (MB), and Bose-Einstein (BE) distributions, plotted as a function of energy E relative to the chemical potential μ_C (or Fermi energy for the Fermi-Dirac distribution).

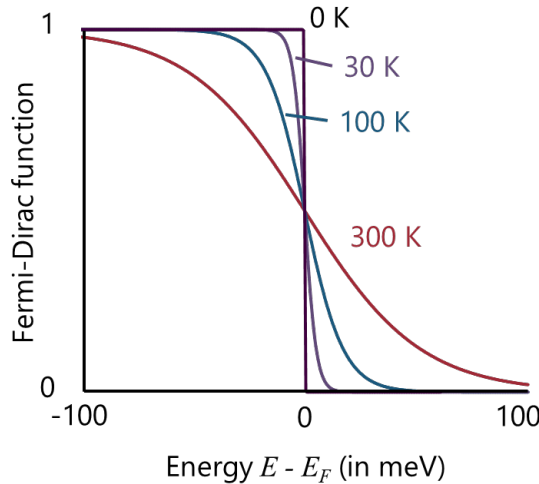


Fig. 5.10. Fermi-Dirac function for temperatures of 0 K, 30 K, 100 K, and 300 K, plotted against the energy relative to the Fermi energy, in meV.

Conventionally, we use the normal definition of temperature, and the chemical potential for electrons is nearly always referred to as the Fermi energy (or Fermi level)³⁷, notated as E_F . We should note that the Fermi energy (or Fermi level) and the chemical potential are *exactly* the same concept for electrons, that is

$$E_F \equiv \mu_C \quad (5.86)$$

³⁷ The terms “Fermi level” and “Fermi energy” are used interchangeably, they have the same meaning, and they both also mean the chemical potential for fermions.

With these conventions, the Fermi-Dirac distribution for electrons and an electron state of energy E is commonly written as

$$f_{FD}(E) = \frac{1}{\exp\left(\frac{E - E_F}{k_B T}\right) + 1} \quad (5.87)$$

This relation is used extensively in the physics of semiconductors, for example, forming the basis for the calculations for many electronic and optical devices.

We see in Fig. 5.9 and Fig. 5.10 several characteristic features of the Fermi-Dirac distribution. For states of energies significantly below the Fermi level, the probability of occupation of any such state approaches 1. (It cannot exceed 1, of course, because we cannot have more than one electron in a state by Pauli exclusion.) Conversely, for states of energies much larger than the Fermi level, the probability of occupation approaches zero. States of energies within a few $k_B T$ of the Fermi level have intermediate probabilities of being occupied. For a state of energy exactly equal to the Fermi level, the probability of occupation is $\frac{1}{2}$, which is sometimes used as a practical definition of the Fermi level.

Note from Fig. 5.10 that, at or very close to absolute zero, any states below the Fermi level are essentially all occupied and any states above the Fermi level are essentially all empty³⁸. With increasing temperature, the curve “softens”, acquiring a transition region round about the Fermi energy or chemical potential, but always going through the point $\frac{1}{2}$ at the energy corresponding to the Fermi energy or chemical potential. In general, we expect some moderate probability of finding electrons in states within a few $k_B T$ above the Fermi level, and we similarly expect that within a few $k_B T$ below the Fermi level, there is some moderate probability that such states are not occupied by electrons.

Bose-Einstein distribution

We could follow a similar argument, starting with the Gibbs factor, to derive the thermal distribution for bosons. In this case, the partition function – the summation over the relative probabilities for the various possible states – is more complicated because we are not restricted to just $N = 0$ or $N = 1$; we can have any number of bosons in a given boson “mode”. We will not give that derivation here (though it is not particularly hard to derive it). The result, which is known as the Bose-Einstein distribution, is, however,

³⁸ It is not quite correct to define the Fermi level as that energy such that, at zero temperature, all states below it are occupied and all states above it are empty; though that is correct if the temperature is indeed zero or close to it, as the temperature changes, even for the same total number of particles in the system, the Fermi level will in general change at least slightly with temperature. This subtlety arises if the distribution of possible states is not uniform in energy. For example, suppose there is a very large number of states just above the energy of the Fermi level at zero temperature. Those states would all be unoccupied at zero temperature, of course. But as the temperature is increased, they would start to become occupied, and the Fermi level would tend to move down so as required to keep the total number of particles in the system constant.

simple to state. The expected number of bosons in a boson mode of energy E per boson is

$$f_{BE}(E) = \frac{1}{\exp\left(\frac{E - \mu_C}{k_B T}\right) - 1} \quad (5.88)$$

Unlike the use of the term “Fermi energy” for the chemical potential for fermions, there is no particular other name for the chemical potential in the case of bosons. This distribution is also plotted in Fig. 5.9.

Note that, though we can think of the Fermi-Dirac distribution as giving the probability that a given state can be occupied, for the Bose-Einstein distribution, because f_{BE} can certainly be greater than 1, we have to think of f_{BE} as being the expected number of bosons in the boson mode rather than the probability the mode is occupied³⁹.

Planck distribution

Compared to the Fermi-Dirac distribution, the Bose-Einstein distribution is used relatively infrequently in its full form as stated here with the chemical potential. That is because most of the bosons we have to deal with in applications, especially in devices, are photons – the quanta of light – or phonons – the quanta of vibrations – or other such quanta associated with oscillations of some kind.

Such bosons are particularly simple in that there is only one possible state for a given such boson in a given “mode”; unlike, say, an atom, there are no excited states of these simple bosons. As a result, the exchange of energy between the system and the reservoir and the exchange of particles between the system and the reservoir are the same physical process – if we know how many particles in a given mode we have exchanged, we know what energy we have exchanged, and *vice versa*. A consequence of that identity of processes is that there is no need for two separate parameters, temperature and chemical potential, to be matched to reach equilibrium; we only need one of them, and by convention we use temperature.

Particles like photons and phonons therefore obey a simpler version of the Bose-Einstein distribution called the Planck distribution. We usually write the energy per boson for these simple bosons in the form

$$E = \hbar\omega \quad (5.89)$$

where ω is the angular frequency associated with the boson mode of interest, such as a mode of the electromagnetic field. Indeed, Eq. (5.89) is the standard expression for photon energy. The Planck distribution can then be stated as

³⁹ Also, the Bose-Einstein distribution has no meaning for energy less than the chemical potential, since that would give a negative number for the result.

$$f_p(\hbar\omega) = \frac{1}{\exp\left(\frac{\hbar\omega}{k_B T}\right) - 1} \quad (5.90)$$

and can be considered as giving the number of such bosons per mode.

It is relatively straightforward to derive Planck's distribution, and we can do that now. Planck's original proposal was that matter emitted light of frequency ν in quanta of size $h\nu$. Subsequently, Einstein proposed that light of a given frequency existed only in such quanta. At the time that was a radical proposal because light was believed to be an electromagnetic wave, and such quantization would seem to be in conflict with the apparent continuous nature of waves.

Part of the resolution of this conflict lies in the notion that the photons exist in modes of the light field, and those modes have the same functional forms as the modes we would calculate classically for electromagnetic waves in resonators or in propagating wave modes. Our modern way of looking at this is to say that a given mode of the electromagnetic field can exist in different levels of excitation, with those levels separated by an energy $E = h\nu \equiv \hbar\omega$.

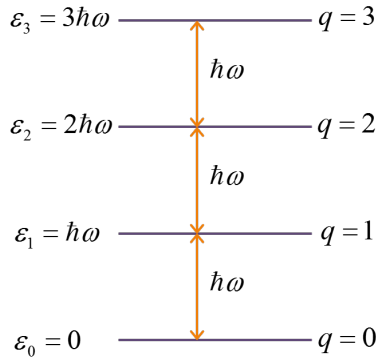


Fig. 5.11. Illustration of the possible energy levels for a mode of the electromagnetic field for which the photon energy is $\hbar\omega$.

The possible energy levels for an electromagnetic mode are sketched in Fig. 5.11 for the first few such levels. We can identify the integer q with the number of photons in the mode if we like, and the energy of such a mode can be written as

$$\varepsilon_q = qh\nu \equiv q\hbar\omega \quad (5.91)$$

where q is zero or a positive integer. It actually makes no difference if we regard q as being the “level of excitation” of the mode or as being the number of photons in the mode; these are just different forms of words for the same underlying physics.

We could add a constant to all the energies ε_q if we wanted⁴⁰; that would make no difference again to the physics here because we will only be concerned with energy differences. For simplicity, we use the form in Eq. (5.91) here.

We presume that, in equilibrium with a reservoir at temperature T , the relative occupation probability of a state of energy ε_q compared to relative occupation probability of a state of energy 0 is given by the Boltzmann factor $\exp(-\varepsilon_q / k_B T)$. To get the absolute probability of finding the oscillator in the state of energy $\varepsilon_q = q\hbar\omega$, we need to normalize the relative probability by dividing by the sum (the partition function) Z of the relative probabilities all the possible states – that is, we divide by the sum

$$Z = \sum_{q=0}^{\infty} \exp(-q\hbar\omega / k_B T) \quad (5.92)$$

This sum in Eq. (5.92) is just a geometric series – that is, a sum in which each term differs from the previous one by constant factor, here $\exp(-\hbar\omega / k_B T)$. The sum of such a series is a standard result, giving

$$Z = \frac{1}{1 - \exp(-\hbar\omega / k_B T)} \quad (5.93)$$

So the probability of finding the mode in the state with energy ε_q is

$$P(q) = \frac{\exp(-q\hbar\omega / k_B T)}{Z} \quad (5.94)$$

The average value of q for this mode in thermal equilibrium at temperature T is therefore

$$\langle q \rangle = \sum_{q=0}^{\infty} qP(q) = \frac{1}{Z} \sum_{q=0}^{\infty} q \exp(-q\hbar\omega / k_B T) \quad (5.95)$$

Now we can use a mathematical trick, noting that

$$\sum_{q=0}^{\infty} q \exp(-qy) = -\frac{d}{dy} \left[\sum_{q=0}^{\infty} \exp(-qy) \right] = -\frac{d}{dy} \left[\frac{1}{1 - \exp(-y)} \right] = \frac{\exp(-y)}{[1 - \exp(-y)]^2} \quad (5.96)$$

So

$$\langle q \rangle = \frac{1}{Z} \frac{\exp(-\hbar\omega / k_B T)}{[1 - \exp(-\hbar\omega / k_B T)]^2} = \frac{\exp(-\hbar\omega / k_B T)}{1 - \exp(-\hbar\omega / k_B T)} \quad (5.97)$$

⁴⁰ In part because we can derive the formal quantum mechanics of light by considering each mode of the electromagnetic field as being a kind of harmonic oscillator, it is common to add $\hbar\omega / 2$ to the energies in Eq. (5.91) because then the energy zero corresponds with the zero of “potential energy” in the oscillator, but whether or not we add this constant or any other makes no difference here to the final result.

So, rearranging Eq. (5.97), the average number of photons per mode in thermal equilibrium at temperature T is the Planck distribution

$$\langle q \rangle = \frac{1}{\exp(\hbar\omega / k_B T) - 1} \quad (5.98)$$

which we see is exactly the same as Eq. (5.90). We see, as proposed, that the Planck distribution is a Bose-Einstein distribution with chemical potential of zero. Note incidentally that $\langle q \rangle$ need not be an integer since it is an average number.

This expression Eq. (5.98) is not yet the spectrum of the emitted light from an object, but we have completed a major step towards that result, and we return to that topic in Chapter 7. The Planck distribution is quite widely useful for many different sorts of vibrating systems, including also vibrations in solids, where the associated quanta are called phonons.

The Maxwell-Boltzmann distribution

We note that, for both the Fermi-Dirac distribution and the Bose-Einstein distribution, as we consider energies very much larger than the chemical potential μ_C (or the Fermi energy), i.e.,

$$E - \mu_C \gg k_B T \quad (5.99)$$

both of them tend to the same mathematical limit, which we can write as

$$f_{MB}(E) = A \exp\left(-\frac{E}{k_B T}\right) \quad (5.100)$$

where

$$A = \exp\left(\frac{\mu_C}{k_B T}\right) \quad (5.101)$$

This distribution, which is also plotted in Fig. 5.9, was originally derived for classical particles, which are presumed to be non-identical⁴¹, and is known as the Maxwell-Boltzmann distribution⁴². In practice with the Maxwell-Boltzmann distribution, A is just regarded as a number that is chosen so as to give the correct total number of particles in the system, though it can formally be related to chemical potential for either fermions or bosons as given by this expression.

⁴¹ It is actually quite hard work to derive it because of this presumption of non-identity, and it is arguably easier to write it down as a limit of the Fermi-Dirac and Bose-Einstein distributions that are based on identical particles. We could even take the view that there are no “non-identical” particles out there anyway, just different species of identical particles, so maybe the Fermi-Dirac and Bose-Einstein distributions are the only correct ones.

⁴² We can see an obvious simple apparent relation between this Maxwell-Boltzmann distribution and the Boltzmann factor, though the simplicity of this end result is somewhat deceptive since the full derivation of the distribution is more complicated than this relation might suggest.

We tend to use this Maxwell-Boltzmann limit wherever we can reasonably get away with it because it leads to much simpler mathematics. With the full Fermi-Dirac distribution it is often much harder to integrate over energy - for example, when we are trying to count the total number of particles in the system.

At high energies, the Fermi-Dirac and Bose-Einstein distribution can tend to this Maxwell-Boltzmann distribution for non-identical particles because this limit applies when probabilities of finding such a high energy state occupied are very small. In that case, because the probability of occupation is much less than 1, the issue of whether we can or cannot have multiple different particles in a state practically does not arise. So, the differences between the counting of fermion states, with their Pauli exclusion principle, boson states, and the states of non-identical particles make negligible difference to the end result.

If we compare the Fermi-Dirac, Maxwell-Boltzmann, and Bose-Einstein distributions in Fig. 5.9, we see that the Bose-Einstein distribution lies above the Maxwell-Boltzmann distribution. We can rationalize this by noting that identical bosons are more likely to be found in the same mode than are classical or non-identical particles; this agrees with the counting we found when comparing (i) the number of possible states of two dollar bills in two boxes, where 2 out of 4 possible states had the two bills in the same boxes, and (ii) the number of possible states of two dollars in two bank accounts, where 2 out of 3 states had the two dollars in the same bank account (there being only one way of having one dollar in one bank account and one in the other). Of course, both the Maxwell-Boltzmann and the Bose-Einstein lie above the Fermi-Dirac distribution because for that distribution there is no possibility of finding two fermions in the one state.

Quiz questions

- Q5.8.1 We have a material that has many possible states for electrons, with many possible corresponding energies for the electron. A small piece of this material is connected to another large piece of material with which it can exchange electrons and also energy. These two pieces of material are at thermal and diffusive equilibrium. Suppose the number of electrons in the system is such that chemical potential corresponds to a Fermi energy of 50 meV and that we are at room temperature, which we take to correspond to $k_{BT} = 25meV$.
- a) The occupation probability for an electron state of energy 50 meV in the small piece of material is $\frac{1}{2}$. True or False?
 - b) Electron states of energy less than 50 meV in the small piece of material have an occupation probability $< \frac{1}{2}$. True or False?
 - c) The occupation probability for an electron state of energy 75 meV is exactly $\frac{1}{3}$. True or False?
 - d) Supposing that the Fermi level remains fixed as we change temperature, then decreasing the temperature of the whole system (i.e., of both pieces of material together) leads to an increase in the occupation probability for all states of energy less than the Fermi level. True or False?

Q5.8.2 Consider a mode for photons of energy E_{ph} (we will take various specific values of this below). Presuming each of the following temperatures T , approximately how many photons are there per mode (presuming the mode is in thermal equilibrium at that temperature)?

- a) At $k_B T = 25 \text{ meV}$ (approximately room temperature) and $E_{ph} = 2 \text{ eV}$ (a visible photon): $\sim e^{80}$, $\sim 1.8 \times 10^{-35}$, or $\sim 2.08 \times 10^{30}$?
- b) At $k_B T = 250 \text{ meV}$ (corresponding approximately to the $\sim 3000 \text{ K}$ temperature of a light bulb filament) and $E_{ph} = 2 \text{ eV}$: 3×10^{-4} , e^8 , or 3×10^4 ?
- c) At $k_B T = 500 \text{ meV}$ (approximately like the sun) and $E_{ph} = 2 \text{ eV}$: 1, 1000, or 0.02?
- d) At $k_B T = 25 \text{ meV}$ (approximately room temperature) and $E_{ph} = 100 \text{ meV}$ (which corresponds to photon energies often used in thermal imagers): 0.02, 1, or 1000?

Q5.8.3 If we raise the chemical potential, then the occupation probabilities for some state of a specific fixed energy E_{sp} rises for all three distributions (Fermi-Dirac, Bose-Einstein, Maxwell-Boltzmann), no matter what is the value of E_{sp} . True or False?

Q5.8.4 If we increase the temperature, then the occupation probabilities for some state of a specific fixed energy E_{sp} rise for all three distributions (Fermi-Dirac, Bose-Einstein, Maxwell-Boltzmann), no matter what is the value of E_{sp} . True or False?

Problems

5.8.1 A system can be in one of four states. These are

- (1) No particles, zero energy
- (2) One particle, energy ε_1
- (3) One particle, energy ε_2
- (4) Two particles, energy $\varepsilon_1 + \varepsilon_2 + I$

where $\varepsilon_2 > \varepsilon_1 > 0$ and $I < 0$. (State (4) might correspond, for example, to a bound state of the two particles.) The system is in thermal and diffusive contact with a large reservoir. All the particles here are the same kind of particle. The temperature of the reservoir is T and its chemical potential for this kind of particle is μ .

- a) Write down the partition function Z of the system. [Hint: the partition function is a sum of the un-normalized or relative probabilities of the various states, where those un-normalized or relative probabilities are given by some appropriate factor. Given that we are working with a system that

can exchange both energy and particles with a reservoir, what factor is appropriate here?]

- b) Give an expression for $N(T, \mu)$, the expected number of particles in the system. [Hint: we can work out actual probabilities P_j for various states j of the system by dividing relative probabilities by the sum of all the relative probabilities. We can work out the average or “expected” value of some quantity by multiplying each possible value of the quantity by its probability, and adding up these results.]
- 5.8.2* In a semiconductor, we find that the probability that the lowest state in the conduction band is occupied by an electron is 10^{-2} at room temperature (300 K). Where is the Fermi level relative to the bottom of the conduction band? Express your answer in meV, and be explicit about the sign of the energy of this Fermi level, stating whether this Fermi level lies in the band gap region (presuming the band gap energy is large) or above the bottom of the conduction band.
- 5.8.3 Consider a metal with Fermi energy $E_F = 5.0$ eV at room temperature $T = 293$ K.
- a) Based on the Fermi-Dirac distribution, what is the probability of finding an electron in a specific state with each of the following energies:
 - (i) $E_1 = 5.025$ eV
 - (ii) $E_2 = 5.050$ eV
 - (iii) $E_3 = 5.075$ eV
 - (iv) $E_4 = 5.0$ eV
 - (v) $E_5 = 4.975$ eV
 - (vi) $E_6 = 4.950$ eV
 - b) Repeat the above calculations, but for a temperature of $T = 77$ K (the temperature of boiling liquid nitrogen). (You may presume the Fermi level still has the same value).
 - c) What is the percentage of error when using the Maxwell-Boltzmann distribution instead of the Fermi-Dirac distribution for (i), (ii) and (iii) for the 293 K case?
 - d) Repeat (c) for the 77 K case.
 - e) Can we use the Maxwell-Boltzmann distribution as an approximation for cases (v) and (vi), for example, for 77 K? Justify your answer.
- 5.8.4 Suppose we have an optical or electromagnetic resonator, which we can consider to be some box with highly reflecting walls. Suppose this resonator is in thermal equilibrium with some heat reservoir. Consider various resonant modes of this box as stated below, and state the average number of photons in the mode at the given reservoir temperatures T in each case.
- a) A mode of wavelength 400 nm (which is blue/violet light) at $T = 5800$ K (the temperature of the Sun)
 - b) A mode of wavelength 10 μm at $T = 5800$ K
 - c) A mode of wavelength 10 μm at $T = 300$ K (room temperature)

- d) A mode of frequency 1 GHz (a typical radio frequency) at $T = 5800$ K
- e) A mode of frequency 1 GHz at $T = 300$ K
- f) A mode of frequency 1 GHz at $T = 2.725$ K (the background temperature of the universe)

5.8.5 In solids, there are many different vibrational modes possible for the crystal. These various modes, known as phonon modes, can be viewed as behaving with the same kind of statistics for particles we call phonons as photons have in electromagnetic modes. One particular type of vibrational modes is called “optical phonons”. If the optical phonon corresponding to a specific optical phonon vibrational mode has an energy of 36 meV, in thermal equilibrium at room temperature, how many phonons are there on average in this mode?

[Notes for interest only (not needed for the question!): Understanding the energy stored in such modes can be useful for many purposes, including calculating the specific heat of materials – essentially how much energy it takes to heat them up. This “optical” name has relatively little to do with optics, in fact. In such modes, adjacent atoms vibrate in opposite directions and some such modes can be excited by shining light on materials in which the adjacent atoms are of different kinds because the different amounts of charge on different adjacent atoms lead to them being pulled in opposite directions.]

5.9 Conclusions

In this chapter, we started out with the simple example of tossing coins. That example illustrates that nearly all possible sequences we will encounter share a common characteristic, which is that they nearly all have equal numbers of heads and tails. We introduced the concepts of a microstate – the equivalent of a particular sequence of heads and tails – and a macrostate – the group of microstates that all have the same number of heads and tails. We saw, especially with the convenience of a Gaussian approximation for evaluating the numbers of microstates, that as the numbers became large, this tendency to find nearly all microstates within a narrow range of “most-likely” macrostates became overwhelming. Even as we constrained the system to a more physical situation of requiring the microstates to have a specific energy, we find this concept is retained.

We can usefully define the concept of entropy as logarithm of the multiplicity of the system – the number of “accessible” microstates. When we allow systems to share energy, we find the dominant macrostates for the two systems correspond to ones in which the rate of change of entropy with energy is the same in both systems, and we identify that quantity with $1/\text{temperature}$. We see that heat flow from hot to cold corresponds to an increase in entropy, and that in practical cases, it is essentially impossible for heat to flow from cold to hot simply because there are massively more accessible outcomes if it flows from hot to cold. The second law of thermodynamics is seen to be essentially the law of increase of entropy, which follows from the observation that systems will tend towards those macrostates of largest multiplicity.

We introduced various useful results for analyzing physical systems, including the Boltzmann and Gibbs factors, and the resulting thermal distributions – Fermi-Dirac, Bose-Einstein (and its simpler Planck version), and Maxwell-Boltzmann. We can now start to look at some of the consequences of these distributions for the behavior of the practical world around us. We will find particularly good examples of their use in electronic and optoelectronic devices and in the statistical mechanics and thermodynamics of light.

Chapter 6

Bands and electronic devices

6.1 Introduction

Modern electronics exploits the different capabilities of three broad classes of materials – metals, semiconductors, and insulators. Simply put, the metals conduct the currents to where we want them to go, the insulators stop the currents from going where we do not want them to go, and the semiconductors control the currents, turning them on and off. We discussed the basic band structures and concepts of metals, insulators and semiconductors in Chapter 4, as well as the idea of “holes” – the absences of electrons in the valence band – which is a particularly important concept for semiconductor devices. Now we expand on these ideas to set up the background concepts for understanding electronic and optoelectronic devices.

When used in electronic devices, metals and insulators are often not crystalline materials (though they can be), but semiconductors mostly are. To get the precise control we need with semiconductors, we have to use very pure and perfect starting materials. Impurities – small amounts of other elements – or imperfections (or “defects”) in the crystal lattice can make significant changes in how semiconductor materials conduct electricity. Controlled “doping” (see Chapter 4 for a first discussion), which adds small amounts of such other elements to semiconductors that are otherwise very pure, allows us in practice to add electrons and/or holes into the semiconductor. Such doping is a major part of semiconductor device design. Such impurities or defects have, however, less effect on electrical conduction in metals or in the basic ability of insulators to insulate.

The simple models we use for metals and insulators typically treat them as if they were at least approximately crystalline, though, so we can describe their properties also in terms of bands of states, at least as a first approximation. Though deeper understanding of the behavior of insulating and metallic materials and structures is important in getting the very best performance out of technologies like those used for semiconductor integrated circuits, the details of the band structures themselves are less critical. So here

we concentrate in some greater depth on key ideas in semiconductors. The combination of band structure concepts together with thermal distributions – especially the Fermi-Dirac distribution that applies to electrons and holes – can give us a very strong basis for understanding semiconductor electronic and optoelectronic devices and the important electrical and optical properties of the materials we use to make them.

6.2 Transport of carriers

We saw in section 4.14 that, near to a band minimum or maximum, we expect parabolic relations between E and components of $\mathbf{k}$. The simplest situation to consider is where this parabolic relation is the same in all three directions, which is what we would call an isotropic parabolic “band” (or region of the band structure). Such a model is a good approximation for many bands near minima or maxima¹. Taking the energy at the minimum to be zero for simplicity, such a relation corresponds to

$$E = \frac{\hbar^2 k^2}{2m_{eff}} \quad (6.1)$$

(which is the same as Eq. (4.45)), where the effective mass m_{eff} is parametrizing the curvature of the parabola. This kind of approach works for both electrons and holes².

Transport of electrons

If we think of an electron in some parabolic minimum (or maximum) in the band as having an energy relative to the minimum (or maximum) as in Eq. (6.1) with effective mass m_{eff} , we can rewrite Eq. (6.1) in terms of crystal momentum of magnitude $p = \hbar k$ (see Chapter 4) as

¹ This isotropic parabolic band approach is a good first approximation near zone center for the valence bands in many semiconductors, including Si, Ge, and the III-V materials that crystallize in zinc blende form, and for the lowest conduction band in zinc blende direct gap III-V materials like GaAs or InP. For the lowest conduction bands in Si and Ge, in part because their lowest points do not lie at zone center, the band minima, though approximately parabolic in each direction, have one effective mass in one direction and a different one in the other two. The approach we are taking here can be extended to such cases, though we will not do that here.

² Electrons in conduction band minima have positive effective masses, often considerably lighter than the actual electron rest mass – ~10% of the electron rest mass is a common order of magnitude for the electron effective mass in many direct gap materials, for example. Electrons near maxima in valence bands do actually have negative effective masses – the parabola is curved downwards, not upwards; however, there we typically think in terms of holes, and, for a variety of reasons that we will discuss later, those negative electron masses can be considered as positive hole masses. Hole effective masses can be quite light, though more commonly the highest valence band has a relatively heavy hole effective mass – ~40% of the actual electron rest mass, for example.

$$E = \frac{p_C^2}{2m_{eff}} \quad (6.2)$$

(which is the same as Eq. (4.46)). Then

$$\frac{dE}{dp_C} = \frac{p_C}{m_{eff}} \quad (6.3)$$

Then, thinking of a velocity v_g such that

$$p_C = m_{eff}v_g \quad (6.4)$$

we have for this velocity at which we expect the particle is moving

$$v_g = \frac{dE}{dp_C} \quad (6.5)$$

Technically, the particular velocity v_g we are working with here is called the “group velocity”³.

Suppose now we apply an “external” force F to the moving “particle”, such as the force resulting from an applied electric field. Then the work done in applying the force through a distance dx is

$$dE = Fdx \quad (6.6)$$

The distance dx is equal to the group velocity v_g times the time, dt , for which the force is applied, so we have

$$dE = Fdx = Fv_g dt \quad (6.7)$$

Hence

³ We are performing a rather informal derivation here, though the result is correct. Normally when we work with waves such as light waves or sound waves in free space, we think that the wave is propagating at some velocity $v = f\lambda = \omega/k$, called the “phase velocity”. Here f is the frequency and λ is the wavelength; correspondingly ω is the angular frequency and $k = 2\pi/\lambda$ is the wavevector magnitude. As long as the phase velocity does not depend on frequency (or equivalently on wavelength or wavevector), such a picture works, and some pulse we might make would propagate with this (phase) velocity also. If, however, the phase velocity is different for different frequencies, which happens at least to a small extent for most wave propagation inside matter, the actual propagation velocity of a pulse is better described by the “group velocity”, which is $v_g = d\omega/dk$. In quantum mechanics, once we consider time dependence, we can generally associate an energy E with an angular frequency ω through $E = \hbar\omega$. In that case, we can more rigorously derive Eq. (6.5) and Eq. (6.9) without even assuming we are at some parabolic minimum. The actual movement of an electron is better considered in terms of some “pulse” or “wave packet”, which has to be made up out of some range of k values, and that wave packet does indeed move at this group velocity.

$$F = \frac{1}{v_g} \frac{dE}{dt} = \frac{1}{v_g} \frac{dE}{dp_C} \frac{dp_C}{dt} \quad (6.8)$$

or, making use of Eq. (6.5)

$$F = \frac{dp_C}{dt} \equiv \frac{d(\hbar k)}{dt} \quad (6.9)$$

So we are saying that, if nothing else changes the crystal momentum, applying a force to an electron, such as with an electric field, leads to

“force is equal to rate of change of (crystal) momentum”

so again this crystal momentum is behaving as if it was the actual momentum of some effective particle of mass m_{eff} . In this picture, then, applying a force moves the electron steadily also through the Brillouin zone.

This kind of transport, in which the electron is effectively continuously accelerated by the applied field, is called “ballistic transport”. It does exist in materials like semiconductors, and is part of the analysis of modern high-speed electronic processes in small devices. Generally, this is only a good approximation for very short distances or short time-scales, however.

More typically, the electron is accelerated (“ballistically”) like this for some average or effective “scattering” time t_s , and then the electron is scattered by some collision. Typical collisions are with other electrons, with crystal vibrations (or “phonons”), or with impurities or defects in the crystal. One simple model is to say that these scattering events are random but strong so that, on the average, the electron velocity is completely randomized by such collisions. That means that the average velocity of the electron after the collision is zero, because it could just as well be going in any direction.

We can be specific now that the applied force is from an electric field of magnitude $\mathcal{E}$, so the magnitude of the force on an electron is

$$F = e\mathcal{E} \quad (6.10)$$

where e is the magnitude of the electron charge. If we accelerate (from zero velocity) using such a force for a time t_s , then the resulting peak momentum just before scattering will be, from Eq. (6.9) with Eq. (6.4) and (6.10)

$$p_C = \frac{dp_C}{dt} t_s = F t_s = e\mathcal{E} t_s = m_{eff} v_{peak}$$

The average velocity of the electron is half this peak value, since we presume it starts on the average from zero velocity and it accelerates linearly to this peak value on the average. So on this model the average value of the electron velocity in the presence of this accelerating field $\mathcal{E}$ is

$$v_{av} = \frac{et_s}{2m_{eff}} \mathcal{E} \quad (6.11)$$

We see in this simple model that the average velocity of the electron is proportional to the applied electric field $\mathcal{E}$. Such transport of an electron is called “drift transport”, and the v_{av} is called the “drift velocity”. Such transport is often written as

$$v_{av} = \mu_e \mathcal{E} \quad (6.12)$$

where

$$\mu_e = \frac{e t_s}{2m_{eff}} \quad (6.13)$$

is called the “mobility”. This kind of “drift” transport, where the electron average velocity is proportional to the applied electric field, will give rise to behavior like Ohm’s law, in which the current (which is proportional to electron velocity) through some “resistor” is proportional to the voltage (and hence the electric field) applied across it.

Transport of holes

At a maximum at the top of the valence band, we see that necessarily the effective mass we would calculate would be negative. This leads to the apparently counter-intuitive conclusion that the electrons would go backwards if we pushed them⁴. Though counter-intuitive, this conclusion is in fact correct. The group velocity of an electron near such a maximum can indeed be “backwards” from what we would expect. In some applied electric field, the electrons are going in whatever direction the group velocity says they should be going in response to the force from an electric field, even if, counter-intuitively, that is the opposite direction we would naively expect.

In terms of current, however, a negative charge going backwards is the same as a positive charge going forwards. So one way to look at this is to pretend that the absence of an electron in some state in such a “negative electron effective mass” maximum is behaving like a positive charge that is indeed going in the same direction as such a positive charge would go in the same electric field. So, though it might seem obvious that the hole, having an effective positive charge, should go in such a direction, this is the result of the cancellation of two negative signs; the actual motion is of negatively charged electrons with negative effective mass going “backwards”. The “absence of an electron” travels in the same direction as the electrons occupying the states round about it in the band structure.

If this seems confusing⁵, the reader can take heart that this model, with its cancelling minus signs, works so well that in practice for devices we can essentially always pretend that the holes behave as positive charges with positive effective masses. We can also

⁴ Remember that when we say “electrons” here we are speaking loosely; these are actually effective “quasi-particles”, with negative charge, that arise from the solution of the Schrödinger equation in a periodic potential; we are not changing the actual mass of “real” electrons or making it negative.

⁵ It seems confusing because it is confusing! Device texts are liable to gloss over this annoying point!

think of the hole kinetic energy as being positive if we look at the band diagram upside down, just as we will be able similarly to think of positive hole energies in the hole Fermi-Dirac distribution when looking at the band diagram upside down.

We can similarly analyze hole transport with ballistic and drift models, just like the electron case, and the “positive hole mass and positive hole charge” effective model continues to work here also.

Quiz questions

- Q6.2.1 In ballistic transport, the electron velocity is proportional to the applied electric field. True or False?
- Q6.2.2 In drift transport, the resulting electrical current is approximately proportional to the applied voltage.
- Q6.2.3 In drift transport, a shorter scattering time (for the same effective mass) leads to a larger mobility.
- Q6.2.4 In drift transport, a lighter effective mass (for the same scattering time) leads to a larger mobility.
- Q6.2.5 Drift rather than ballistic transport is the dominant transport mechanism in a normal resistor (i.e., one obeying Ohm’s law).
- Q6.2.6 When we apply a voltage to a piece of semiconductor, holes in a valence band with an approximately parabolic maximum in energy as function of wavevector magnitude k , will give an electrical current in the same direction as that of electrons in a parabolic minimum in the conduction band.

Problems

- 6.2.1 a) If a piece of GaAs is doped with beryllium (Be), with the Be atoms taking the places of some of the Ga atoms, is it p-type or n-type? (Justify your answer briefly.)
- b) If silicon atoms are added to a piece of GaAs and they occupy gallium sites, do silicon atoms work as acceptors or donors? (Justify your answer briefly.)

6.3 Fermi levels and doping

Electrons and holes in semiconductors exist essentially always in thermal distributions⁶. Those thermal distribution strongly influence how semiconductor devices work. To

⁶ Even if we inject or remove electrons or holes in a semiconductor, whatever population there is of either of them essentially instantaneously (e.g., in picoseconds or less) settles into a thermal distribution, at least locally. The reason is that the scattering of electrons or holes off one another is a very strong process (because of the Coulomb interaction) that very quickly randomizes the energies of the electrons or holes into thermal distributions. The temperature of that distribution

examine such thermal populations in semiconductors and insulators in more detail, we need to use our thermal distribution functions and our knowledge of the energies of possible states from the band structure. Here we can give a qualitative discussion of the key ideas. For a more quantitative discussion, we need to formally introduce the idea of “density of states” – the number of states within a given energy range in a band. That is a straightforward concept, but we do not need it for the rest of the qualitative discussion here, so we have put this more detailed discussion in Appendix C. There we combine the details of the Fermi-Dirac distribution and this density of states to give some useful algebraic results that formally allow us to relate Fermi energy (also known as the Fermi level) and carrier densities. After reading the qualitative discussion here, students interested in electronic and optoelectronic device physics can proceed to this Appendix.

We saw in Chapter 5 that electrons in thermal equilibrium should have a Fermi-Dirac distribution – that is, the probability of finding an electron in a state of a given energy E should be given by the Fermi-Dirac factor (Eq. (5.87))

$$f_{FD}(E) = \frac{1}{\exp\left(\frac{E - E_F}{k_B T}\right) + 1} \quad (6.14)$$

We can see that, if we knew the Fermi energy E_F and the energies E of all the states that the electrons could be in, then we could also calculate the total number N_e of electrons just by adding up the probabilities of occupation of all those states. It is also true that if we knew N_e and the energies of all possible electron states, we could deduce the Fermi energy, though that is not typically a simple calculation with a closed algebraic form, and it usually has to be performed numerically. Nonetheless, in a given material, E_F determines N_e and N_e determines E_F .

We can usefully understand, at least qualitatively, how the position of the Fermi level varies as we dope the material. See Fig. 6.1 and Fig. 6.2.

In a pure semiconductor – so an “undoped” semiconductor with no dopant atoms (also known as an “intrinsic” semiconductor) – we know that the numbers of electrons in the conduction band must be equal to the number of holes in the valence band. Otherwise, we would not have electrical neutrality. As a result, the Fermi energy has to be near the middle of the band gap⁷.

could be different in different parts of the semiconductor. Those “carrier” temperatures can be different from the background crystal (or “lattice”) temperature. The Fermi levels for electrons and holes need not be the same, and those Fermi levels can also vary between different parts of a device structure, but the local distributions will still very likely be thermal distributions, at some temperatures and Fermi levels, because of this strong scattering.

⁷ If the electron and hole effective masses, and hence densities of states, are not equal, it will not be exactly at the middle, but it will be near the middle within some amount of energy of the scale of $k_B T$.

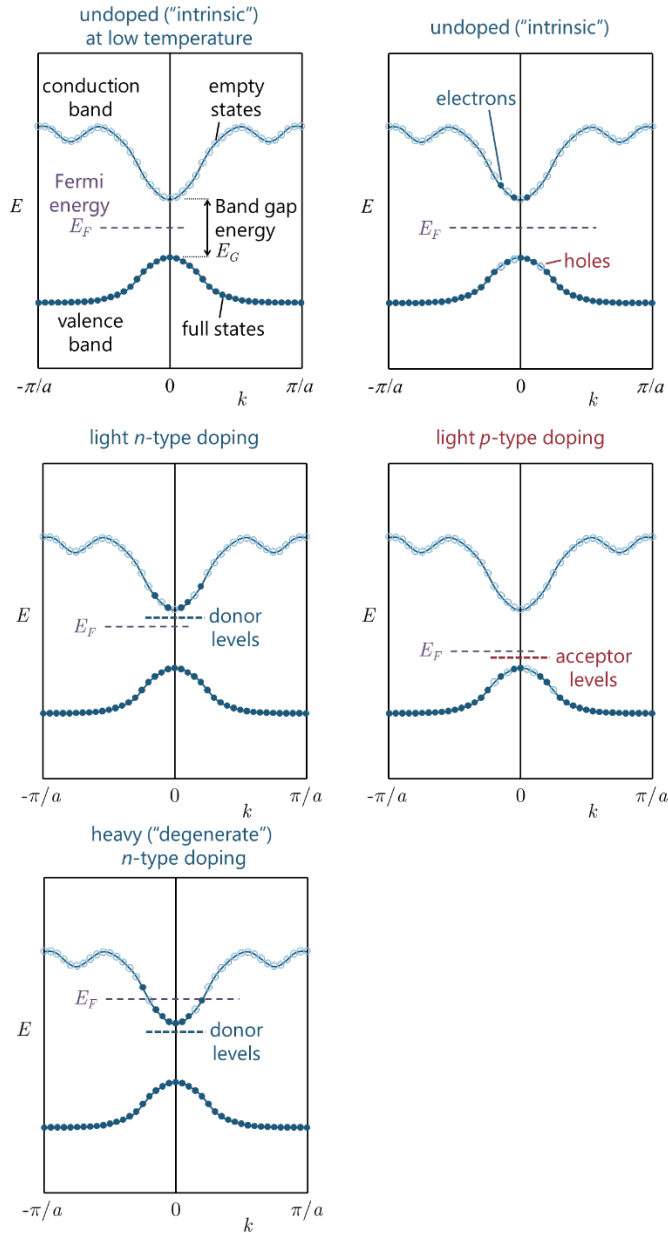


Fig. 6.1. Band diagrams showing electron and hole populations as a function of doping and the resulting Fermi energy for a hypothetical semiconductor, here shown for k space in one direction. (Not to scale.) The middle panels presume relatively light n or p doping, so the Fermi level is below the donor levels or above the acceptor levels. The bottom panel shows heavy “degenerate” n doping, with the Fermi level significantly inside the conduction band.

Semiconductor band gap energies of common semiconductors used for electronic and optoelectronic devices at room temperature or below are mostly in the range from ~ 0.5

eV to 3.5 eV. Even at room temperature, which corresponds roughly to $k_B T \sim 25$ meV, the band edges for both valence and conduction bands are therefore $> 10 k_B T$ away from the Fermi level, so the Fermi-Dirac factor is a small number for valence and conduction band states.

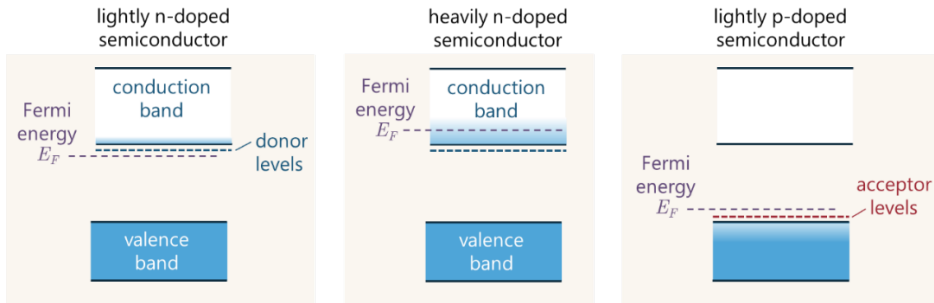


Fig. 6.2. Band diagrams showing electron and hole populations as a function of doping and the resulting Fermi energy for a hypothetical semiconductor, here shown for one direction in real space. (Not to scale.)

At room temperature, though, this can still correspond overall to non-negligible overall electron and hole concentrations in such an intrinsic semiconductor, especially for semiconductors with small band gap energies (“narrow-gap” semiconductors). At cryogenic temperatures, though, even narrow-gap intrinsic semiconductors will have very low overall electron and hole concentrations.

If we dope the semiconductor n -type, we add in “donor” energy levels near the conduction band, and also potentially a free electron in the conduction band, and we shift the Fermi level somewhere near to the conduction band edge. Note that the donor energy levels themselves have an occupation probability given by a Fermi-Dirac factor also, and this can be used, together with the known density of electron states in the conduction band, to deduce the probability of the donor atoms being “ionized” to give their electron to the conduction band. Since the separation of the donor levels from the conduction band edge is typically quite a small number, such as 10 meV, then at low doping densities, the probability of this ionization is quite large. With increasing n -type doping density, the Fermi level will move progressively from near the middle of the band gap up towards the conduction band edge.

When the Fermi level is still several $k_B T$ or more below the conduction band edge, the Fermi-Dirac factor and distribution can be approximated with their mathematically simpler Maxwell-Boltzmann limit. With further increasing doping density, the Fermi level can also move up past the energy of the donor levels, and even into the conduction band. If the Fermi level has moved a few $k_B T$ or higher into the conduction band, we call this situation “degenerate” doping, or a “degenerate” semiconductor. In all such cases where the Fermi level is closer than a few $k_B T$ from the band edge or when it has moved above the band edge into the fully “degenerate” situation, the full Fermi-Dirac factor or distribution has to be used to describe the statistics accurately.

Similar phenomena happen with increasing p -doping, but in the opposite sense. Adding p -doping to a pure (intrinsic) semiconductor will progressively move the Fermi level from near the middle of the band gap energy down towards the valence band⁸.

As a practical matter, when we are deliberately doping a semiconductor material either n -type or p -type, the Fermi level will essentially always end up within a few $k_B T$ of the edge of the corresponding band (the conduction band for n -type doping, and the valence band for p -type doping). On the electron-volt-sized scale on which we draw energy band diagrams, this means that graphically the Fermi energy will always be essentially very close to the corresponding band edge for such doped materials.

6.4 Diodes

So far, we have discussed isolated semiconductor materials, considering their Fermi levels separately for each such material. Much of the practical usefulness of semiconductor materials comes when we join them together. One of the simplest and most useful such structures is a p - n diode, where we join⁹ an n -type material and a p -type material. We will look first at what happens when we join the n -doped and p -doped regions, but without applying any voltage to the diode.

Joining n and p regions

We can understand what happens first in a “mechanistic” picture based on the diffusive movement of free charges. Just before conceptually “joining” p -doped and n -doped semiconductors, the p -doped semiconductor has ionized acceptors and therefore a number of “free” holes in the valence band equal to that number of ionized acceptors; similarly the n -doped semiconductor has ionized donors and a number of “free” electrons in the conduction band equal to that number of ionized donors. Fig. 6.3 illustrates these situations.

Before joining them, therefore, each piece of semiconductor has a number of fixed ionized impurities and the same number of corresponding mobile or “free” carriers¹⁰.

⁸ Degenerate p -type doping is relatively hard to achieve because of the heavy effective mass typically found in the highest valence band. That means that a very large number of holes would be required to move the Fermi level substantially into the valence band.

⁹ In practice, we essentially never just take a piece of n -type material and a piece of p -type material and then somehow stick them together to make such a diode. Usually such structures are grown either with these regions of different doping built in during the growth of the crystal itself by varying the added doping during the progressive growth of the crystal, or the dopants are diffused or implanted into the material after the crystal growth.

¹⁰ Note that, before joining, there is no requirement that the number of ionized donors in the n -type material is the same as the number of ionized acceptors in the p -type material, even if the figure is drawn this way for graphic simplicity. After joining, the numbers of ionized acceptors and ionized donors in the depletion region of the diode will, however, be equal, preserving charge neutrality overall.

In this picture, there are obviously free holes in the p -type material, but almost no free electrons there.¹¹ Similarly, there are free electrons in the n -type material but almost no free holes there¹¹.

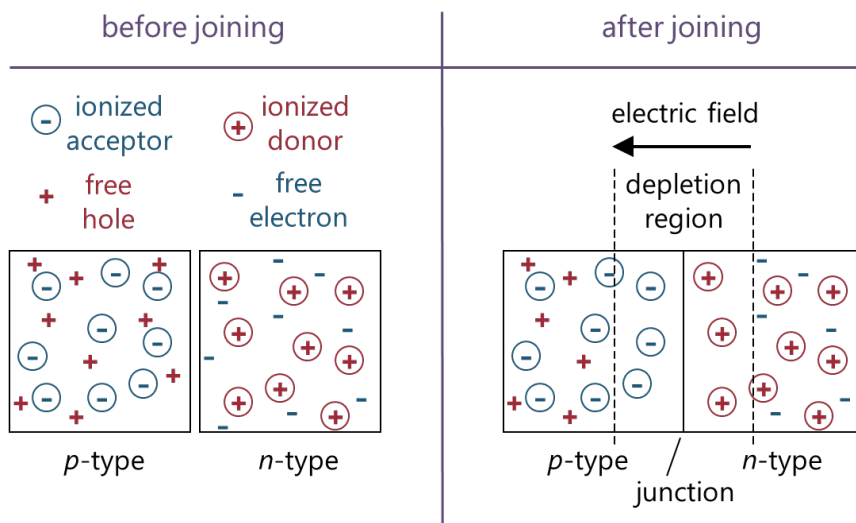


Fig. 6.3. Sketch of p and n doped regions before (left) and after (right) joining.

After we join them, we expect the process of diffusion to take place, for both electrons and holes. “Random walks” of free electrons in arbitrary directions will lead to them on the average moving towards the regions of lower free electron density, simply because there are fewer free electrons from there that might be heading towards regions of higher free electron density; the resulting process of particles tending to move towards regions of lower particle density for a given kind of particle is called diffusion. We can view the same random walk diffusion process as taking place for the free holes. So electrons will initially tend to diffuse from right to left and holes will tend to diffuse from left to right.

As the electrons and holes diffuse, they upset the charge balance in the regions near the junction of the two materials. We can view the electrons and the holes in a given region as essentially canceling one another out; electrons will tend to fill in any “holes” they find, and similarly holes will tend to fill up with any electrons they “find”. So in a region near the junction, there will be relatively few electrons and holes – a region called the depletion region (so-called because it is “depleted” of mobile charges). The fixed charges from the ionized acceptors and donors are still present in this depletion region, and these separated positive and negative charges give rise to an electric field, a field that acts in such a direction as to discourage any further diffusion of the mobile charges by pushing them in the other direction.

¹¹ There is in fact a very small number of electrons in the conduction band of the p -type material, and similarly a very small number of holes in the valence band of the n -type material; we will return to this point when we discuss diode currents below. These numbers are so small that they do not affect the “electrostatic” argument about charges and voltages, but they are important in understanding especially “reverse” leakage current in diodes.

This mechanistic picture is correct, but we can also understand what is happening based on chemical potentials (here, Fermi levels), which leads to a simple and precise characterization of the result of joining these materials. From our discussion of equilibration¹² when we are able to exchange particles, we concluded that the chemical potential must be the same in such connected regions when we are in equilibrium. So, on joining these two materials and allowing them to reach diffusive equilibrium (at the same temperature), then we know that the chemical potential, or here, equivalently, the Fermi energy, must be the same throughout the combined structure.

In the n -type material, we know that, at least far from the junction, the Fermi energy will be somewhere near the conduction band edge in some moderately n -doped material, and in the p -type material it will be somewhere near the valence band edge in some moderately p -doped material, as sketched in the top part of Fig. 6.4. Once we join the materials and allow the electrons and holes to diffuse, then we know that the Fermi level must be the same throughout the structure after we have achieved equilibrium, as sketched in the bottom part of Fig. 6.4.

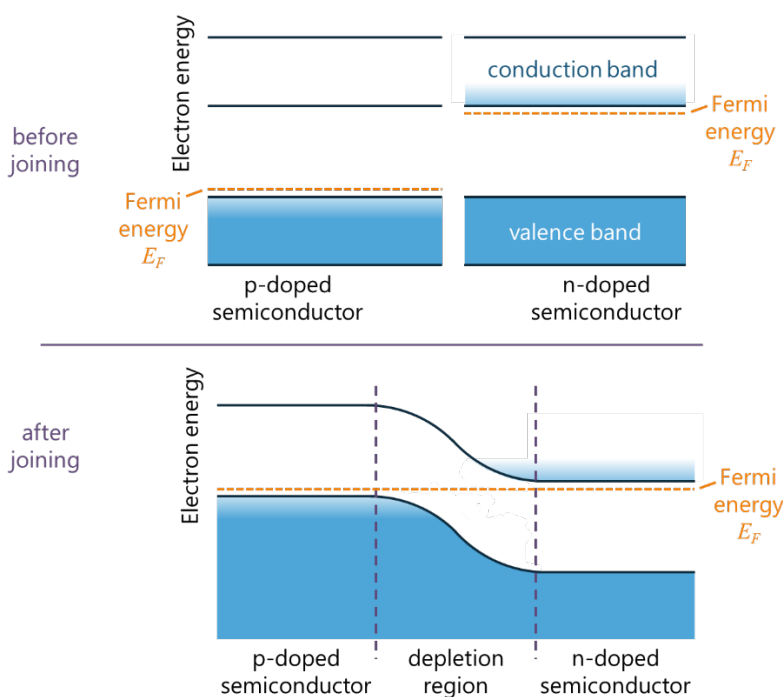


Fig. 6.4. Sketch of the band structure as a function of position before (top) and after (bottom) joining p - and n -type semiconductors.

To a good approximation, in the “depletion region” in the middle of the combined structure, there is little or no free charge; only at the very edges of this region do we get to a slightly ambiguous region where we start to have some amount of free charge in the form of free electrons and holes; that region is generally quite small in electronic

¹² By “equilibration” we simply mean “coming to equilibrium”.

device structures that are at least moderately large, so a simple approximation is just to assume that this depletion region ends essentially abruptly at both sides. Under that “depletion approximation” it is relatively easy to solve for the full behavior of the structure, including the precise form of the “bending” of the band edges inside the depletion region, though we will not do that here.

Applying voltage ¹³

To understand from our statistical physics what happens when we apply voltage to a diode, we need the other equivalent definition of chemical potential that we introduced previously in Eq. (5.71) and which is proved there. In this definition, the chemical potential also corresponds to the amount of energy added per particle as we add particles, provided that in doing so we keep the entropy constant; that is, restating this relation here,

$$\mu_c = \left(\frac{\partial U}{\partial N} \right) \bigg|_S \left(\equiv -\tau \left(\frac{\partial \sigma}{\partial N} \right) \bigg|_U \right) \quad (6.15)$$

So, with this definition, we see that, if we raise the electrostatic potential of a set of charges by raising the voltage, then the chemical potential associated with adding any such particles rises accordingly. Merely changing the electrostatic potential seen by a whole collection of charges in a system by effectively “lifting the system up” in an electrostatic sense does not change the entropy of the system. So, changing the energy per particle of a set of charges by raising the voltage merely corresponds to raising the chemical potential of that set of charges by exactly the same amount of energy per particle.

As a result, if we change the voltage in volts by some amount, we merely change the chemical potential by the particle’s charge times that voltage¹⁴. If we choose to write energies in electron-volts, then the change in energy of an electron when raising the voltage by V volts is simply

$$\Delta E = -e \times V \text{ J} \equiv -V \text{ eV} \quad (6.16)$$

So, in diagrams like Fig. 6.4 or Fig. 6.5, applying a voltage V to some conducting region is equivalent to moving the band diagram and the chemical potential for that conducting region by an amount $-e \times V$ joules or an amount numerically equal to $-V$ electron-volts¹⁵.

¹³ Note this online lecture includes the material in Chapter 5, section 5.7, deriving Eq. (6.15).

¹⁴ Remember that voltage is joules/coulomb. The potential energy change of a particle of charge Q coulombs on raising its electrostatic potential by an amount V volts is simply QV .

¹⁵ Since we are applying a positive voltage to the p material in Fig. 6.5, because of the “-” sign here we are therefore moving the p region down on this diagram (relative to the n region).

Biasing diodes 21d

If we apply a “forward” bias to a diode, as illustrated in Fig. 6.5, various things happen. First, we should note that, because of the “drive” from the voltage source, which is continually now putting energy into the system, such a situation is no longer one of thermal or diffusive equilibrium¹⁶. Outside the depletion region, but within the n -doped region on the right, however, the electron distribution will still locally be one in which the material is approximately in thermal equilibrium, and it is meaningful still to define a Fermi energy for the electrons. The same is true outside the depletion region but within the p -doped region on the left. These Fermi levels are, however, now different, with the one on the left having been lowered by the corresponding reduction in electrostatic potential energy for the electrons on the left compared to those on the right because of the applied voltage V .

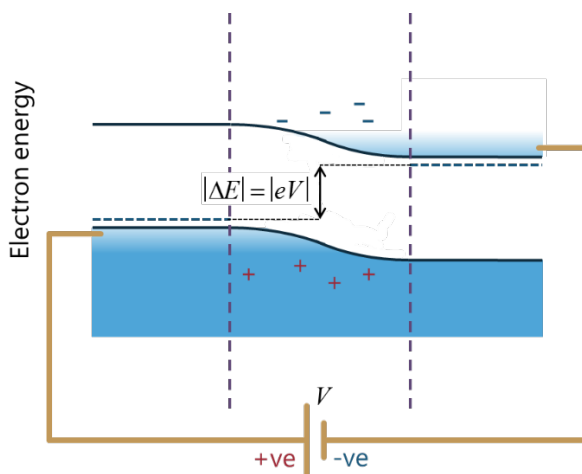


Fig. 6.5. Band diagram for a diode under forward bias with a voltage V from a battery.

Given that the Fermi level for the electrons is lower on the left than it is on the right, electrons will try to diffuse from the right to the left within the conduction band. Similarly, holes will try to diffuse from the left to the right because they see a “lower” Fermi level for holes on the right (remembering we have to “stand on our heads” to visualize what happens for holes). Both of these diffusions added together are what gives the “forward” current in a diode¹⁷.

Once some electrons manage to get from the right to the left, they will then essentially equilibrate to form part of the thermal distribution of electrons on the left, with the corresponding Fermi level on the left. This equilibration process has to involve what

¹⁶ Such a situation that may be stable in time – e.g., with a steady current flowing – but that is not necessarily in equilibrium is called a “steady state” situation.

¹⁷ There is, incidentally, no requirement that the hole current and the electron current are equal in magnitude. High n -doping on one side and low p -doping on the other would lead to large electron current but small hole current.

are known as “recombination” processes, by which an electron can fall from the conduction band to “fill in” a hole in the valence band. Similar processes happen on the right for holes that manage to diffuse from the left to the right.

The situation for a reverse-biased diode is sketched in Fig. 6.6. Here we see that the barrier for the previous diffusion of electrons from the n -doped material into the p -doped material has become very high, and similarly for the diffusion of holes from the p -doped material into the n -doped material, so this “forward” diffusion current becomes exponentially small. Hence, we obtain the basic “rectifying” behavior of a diode, which is that it passes current relatively well in one direction and relatively not well in the other.

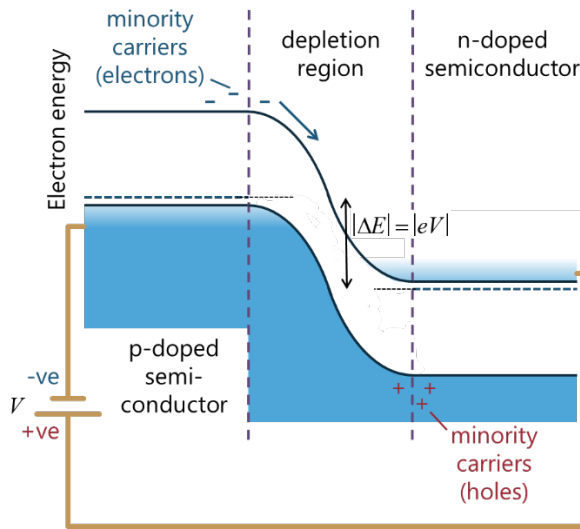


Fig. 6.6. Band diagram for a diode under reverse bias with a voltage V from a battery.

There is, however, a small “reverse” current that flows in the “reverse-biased” diode. In the p -doped material, there is still a relatively small but non-zero electron concentration in the conduction band. Electrons in the conduction band in a p -doped material are known as “minority carriers”; by contrast, holes in a p -doped material are “majority carriers”. There are, of course, very many more majority carriers than there are minority carriers. Our Fermi-Dirac statistics enables us to calculate just how large this number of minority carriers is.

We have indicated this small number of (minority carrier) electrons in the conduction band in the p -doped region just at the edge of the depletion region in Fig. 6.6. (This density of minority carriers extends through the whole p -doped region but, for clarity, we are only sketching them near the edge of the region here.) These electrons see a lower Fermi level to their right in the n -doped material, and so they will tend to diffuse over there¹⁸.

¹⁸ To be more complete, we could also say that the minority carriers in such a reverse-biased diode diffuse into the depletion region, and are then “swept” by the strong electric field there, in

A similar process exists for the small number of holes in the valence band in the n -doped material; these holes in the n -doped material are the minority carriers there (with the electrons in n -doped material being the majority carriers there). These minority carrier holes will tend to diffuse over to the p -doped material. This small current under reverse bias is known as “reverse leakage”, and the sum of these two “minority carrier diffusion currents”, which is the magnitude of this reverse leakage current, we call the “saturation current” I_S . Note that we cannot completely get rid of this reverse current in diodes; diodes cannot be absolutely perfect rectifiers.

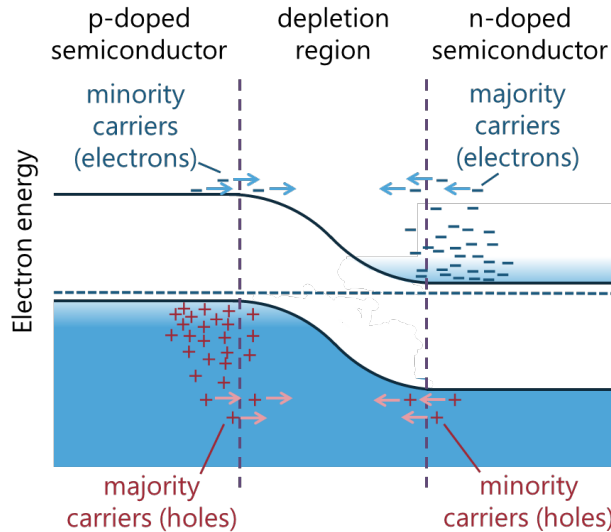


Fig. 6.7. Sketch of balancing diffusion currents between the two sides of a diode at zero bias. Minority and majority carriers are sketched explicitly near to the edges of the depletion region, though these densities extend through the rest of the p -doped region on the left and through the rest of the n -doped region on the right.

Majority carrier electrons diffuse from the n -doped material into the p -doped material and minority carrier electrons diffuse from the p -doped material into the n -doped material. Similarly majority carrier holes diffuse from the p -doped material into the n -doped material and minority carrier holes diffuse from the n -doped material into the p -doped material. At zero bias, all these diffusion currents balance out and we have diffusive equilibrium, consistent with the constant Fermi level throughout the structure, as shown in Fig. 6.4 and Fig. 6.7. So, at zero applied bias (as in Fig. 6.7), the diffusion current from the majority carriers on each side has to balance that from the minority carriers diffusing in the opposite direction, so this majority carrier diffusion will lead to a “forward” current of magnitude I_S also, in a direction opposite to the “reverse leakage” minority carrier diffusion.

what may then be more like “drift” transport, into the other side of the diode. The amount of this current is, however, controlled by the initial diffusion process, so that process sets its magnitude.

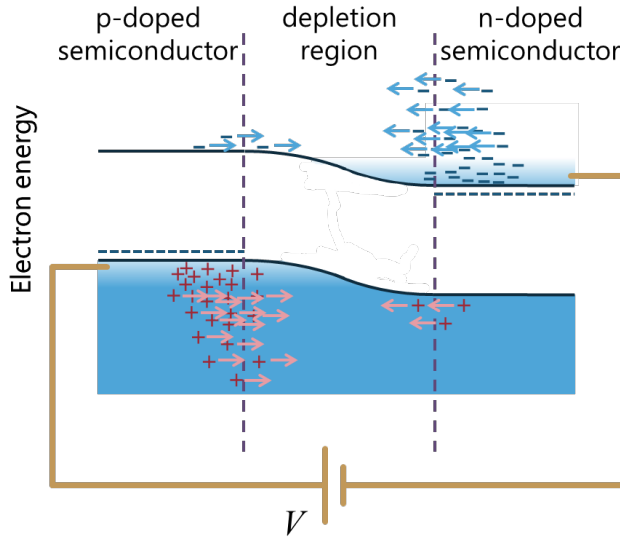


Fig. 6.8. Sketch of the diffusion currents under forward bias. The majority carrier diffusion of both holes from the left and electrons from the right has increased, giving net “forward” current.

We will not give a full derivation of the diode current-voltage characteristic here; to do so would require that we treat the diffusion and recombination processes more quantitatively than we have space to do here. It is, however, now straightforward to write down the overall form of that characteristic in a simple model.

Once we forward-bias the diode, as shown in Fig. 6.8, presuming a Maxwell-Boltzmann approximation for occupation probabilities of electron states, we have increased the occupation probability of all the electron states in the conduction band in the n -doped semiconductor by an amount $\exp(eV / k_B T)$ because we have raised the Fermi level by an amount of magnitude eV ; hence, the majority carrier diffusion current of electrons from the n -doped material increases by this factor. A similar phenomenon happens for the majority hole diffusion from the p -doped material. Since the forward diffusion current for $V = 0$ was of magnitude I_s (to balance the “reverse leakage” minority carrier diffusion), the forward diffusion current is now, after multiplying both electron and hole forward diffusion currents by the same exponential factor, a total of $I_s \exp(eV / k_B T)$. To get the net forward current, we merely now need to subtract the “reverse leakage” minority carrier diffusion current to obtain

$$I = I_s \left[\exp\left(\frac{eV}{k_B T}\right) - 1 \right] \quad (6.17)$$

which is the current-voltage characteristic of an “ideal” diode.

In Eq. (6.17), the “-1” term is the “reverse” diffusion of the minority carriers, which gives the reverse leakage current, and the $\exp(eV / k_B T)$ term is the “forward” majority carrier diffusion. In summary, with no applied voltage, the majority carrier diffusion is balanced by minority carrier diffusion, but with increasing forward bias, the majority

carrier diffusion increases exponentially to give this classic, approximately exponential curve, as sketched in Fig. 6.9.

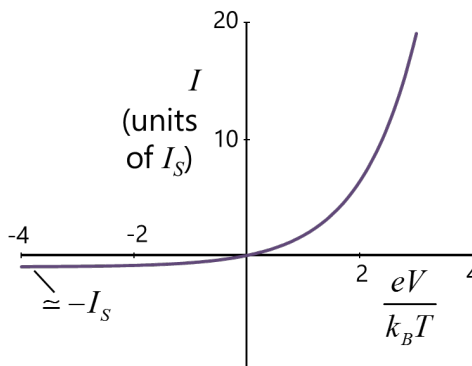


Fig. 6.9. Current (I) vs. Voltage (V) characteristic of a diode, with current in units of I_s , and voltage in units of $k_B T / e$.

Quiz questions

- Q6.4.1 When we join an n -doped piece of semiconductor to a p -doped piece of semiconductor, we should expect holes to diffuse from the p -doped semiconductor to the n -doped semiconductor. True or False?
- Q6.4.2 Near the junction of n -type and p -type semiconductors, we expect to find ionized acceptors inside the n -type material. True or False?
- Q6.4.3 When we join n -type and p -type semiconductors, without making any electrical connections to the semiconductor materials, we expect that there will be an electric field near the junction, in such a direction as to push holes into the p -type material. True or False?
- Q6.4.4 The depletion region in a p - n diode is a region with no acceptor or donor atoms. True or False?
- Q6.4.5 In a diode with no applied voltage, we expect the Fermi levels for electrons to be the same in the n -type and p -type materials. True or False?
- Q6.4.6 Chemical potential can also be defined as the rate of change of energy with the number of particles, at constant entropy. True or False?
- Q6.4.7 Applying a voltage V between the two sides of a diode separates the Fermi levels on the n -type and p -type semiconductors by an energy of magnitude $|e \times V|$, where e here is the (magnitude of the) electronic charge. True or False?
- Q6.4.8 Forward-biasing a diode leads to electrons diffusing from the p -type semiconductor to the n -type semiconductor. True or False?
- Q6.4.9 Electrons are majority carriers in n -type semiconductors. True or False?
- Q6.4.10 Holes are minority carriers in n -type semiconductors. True or False?

- Q6.4.11 Reverse leakage current is caused by majority carriers diffusing. True or False?
- Q6.4.12 The forward diffusion current in a diode is caused by minority carriers diffusing. True or False?
- Q6.4.13 In an ideal diode, the reverse leakage current at a high reverse bias will be approximately equal in magnitude to the forward current at a forward bias voltage V such that $eV / k_B T = \log 2$. True or False?

Problems

- 6.4.1 In a diode with no bias applied to it, the total number of electrons in the conduction band of the n -type material has to be the same as the total number of holes in the valence band of the p -type material. True or False? (Justify your answer briefly.)
- 6.4.2 Electrons are the minority carriers in a p -doped material. True or False? (Justify your answer briefly.)
- 6.4.3 In a diode at zero bias at equilibrium, the Fermi level used to describe the electron distribution in the p -doped material is at the same energy as the Fermi level used to describe the electron distribution in the n -type material. True or False? (Justify your answer briefly.)
- 6.4.4* Consider a specific ideal diode at room temperature (which for this question we can take to correspond to $k_B T = 25$ meV). At a large reverse bias (so at some reverse voltage of magnitude $|V| \gg 25$ mV), the current flowing in this diode has a magnitude of 1 nA.

Suppose now we apply various other voltages to this diode. For each of the following voltages, what is the magnitude of the current flowing through the diode? For any non-zero answers, state also whether this current is in the “forward” direction or the “backward” direction (where the “backward” direction corresponds to the direction of current flow under large reverse bias.. [Note: a positive voltage here corresponds to a forward-bias voltage.]

- (i) -25 mV
 - (ii) 0 V
 - (iii) 25 mV
 - (iv) 100 mV
 - (v) 500 mV
- 6.4.5 We want to use a semiconductor diode as a temperature sensor. We presume the diode can be regarded as an “ideal” diode. The current through the diode at large reverse bias is of some magnitude I_S . We then forward-bias the diode in such a way that, regardless of the temperature or the resulting voltage across the diode, the (forward) current through the diode is always the same, at some value I_B . (Such “constant current” biasing could be accomplished, for example, by

biasing through a very large series resistor from a very large (forward) power supply voltage, or by various other known circuits.)

- (i) Derive an expression for the voltage across the diode as a function of temperature T (in kelvin).
- (ii) Suppose for this ideal diode that $I_S = 1 \text{ nA}$ and that the (forward) bias current is set at $I_B = 1 \text{ }\mu\text{A}$. Presume the temperature changes from room temperature of $20 \text{ }^\circ\text{C}$ (293.15 K) to $30 \text{ }^\circ\text{C}$. How much does the voltage across the diode change?

6.5 Conclusions

In this chapter, we have seen how the concepts of bands of states and of thermal distributions allow us to understand much of the basic operation of crystalline semiconductor devices. There are, of course, many other aspects about the physics of such devices that we have not discussed here, but these core ideas we have introduced give us many of the concepts and vocabulary that allow us to understand a broad range of mechanisms and devices in electronics and optoelectronics.

Chapter 7

Light and quantum mechanics

7.1 Introduction

The history of the quantum mechanics of light goes back to the very start of quantum mechanics itself, with Planck's proposal that light is emitted in quanta of size $E = h\nu \equiv \hbar\omega$. Famously, this proposal led to the correct explanation of the spectrum of light emitted by hot bodies, which in its ideal form is known as the black-body spectrum. Right from the beginning, then, light brings together the two main topics of this book – quantum mechanics and thermal physics of statistical mechanics and thermal distributions.

We now have enough understanding of both the quantum and thermal aspects to derive and explain the main behaviors of light and how it interacts with matter. We can explain light emission and absorption, and the relation between them, in everyday bodies; we can understand key practical issues such as the limitations of conventional light bulbs and the light from the sun; and we can introduce key modern concepts such as the process of stimulated emission that makes lasers work.

We already derived the thermal distribution – Planck's distribution – that lies behind the statistical mechanics and thermodynamics of light in Chapter 5. We will go on to derive key concepts in the thermodynamics of light, including black-body spectra and Kirchhoff's law of radiation, and we will derive Einstein's famous "A and B coefficient" argument that relates emission and absorption processes, including stimulated emission. We will start with the photoelectric effect, whose explanation by Einstein in 1905 was a key step in the development of quantum mechanics in general and of light in particular.

7.2 The photoelectric effect

In 1887, Heinrich Hertz had observed that if you took some metal plates, put a voltage between them, and then shone ultraviolet light on one of them, the spark that he could get between the plates became brighter. In 1902, Philipp Lenard, who had been Hertz's assistant, continued these experiments, and was able to clarify that there was flow of negative charge between plates in a vacuum when they were illuminated, as shown in Fig. 7.1.

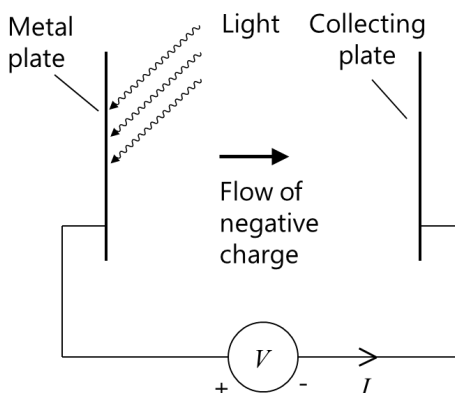


Fig. 7.1. Apparatus for observing the photoelectric effect.

He also found that, for a given wavelength of light, he could stop the current altogether with a specific negative bias voltage on the collecting plate. We can interpret that as giving us a maximum velocity with which the electrons are emitted from the metal plate, with the force from the electric field between the plates over the distance between the plates being just enough to stop all the emitted electrons.

Equivalently, we can deduce the maximum kinetic energy with which the electrons are emitted from the plate in the direction between the plates, with that kinetic energy corresponding to the potential difference between the plates when we apply just enough potential – the “stopping voltage” to stop the flow. (That kinetic energy in electron-volts would be numerically equal to the magnitude of the voltage between the plates.)

The surprise here is that, with an obvious classical way of looking at the effect, we would think that shining a brighter light on the plate would lead to electrons with more kinetic energy. For example, if we were heating the plate up with the light, we might expect that, with a brighter light, the emitted electrons would have more kinetic energy because there was more power in the light beam. But this is not what happens (and anyway, in such experiments, we are not usually shining enough power on the plates to heat them up to any substantial extent).

Instead, though we get more electrons emitted with more light, the “stopping voltage” is the same – that is, the electrons always seem to be emitted with the same maximum kinetic energy. And, furthermore, if we change to a shorter wavelength, we have to increase the “stopping voltage” to stop the flow of current. With shorter wavelengths,

the kinetic energy of the electrons is therefore apparently greater, independent of the brightness of the light.

The explanation of the photoelectric effect proposed by Einstein is as shown in Fig. 7.2. We presume that electrons in metal sit in some sort of “pool”, which we now understand to be a Fermi-Dirac distribution, filled up approximately to the Fermi level. There is an energy barrier, called the work function, ϕ , here, that electrons have to overcome to get out of the metal. A photon of energy $h\nu$, exceeding the height of the barrier, will correspondingly raise the energy of the electron, allowing it to be extracted from the metal. The excess energy the electron has from the photon will appear as kinetic energy of the electron.

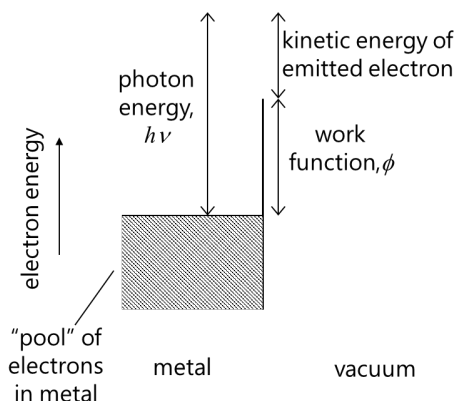


Fig. 7.2. Explanation of the photoelectric effect. Light is composed of photons of energy $h\nu$, where ν is the frequency. The kinetic energy of the emitted electron is given by the difference between the photon energy and the work function, ϕ .

In the worst case of the electron traveling directly away from the metal plate towards the collecting electrode, all of that kinetic energy will correspond to motion in the direction towards the collecting electrode, and we will need the full “stopping voltage” to slow that electron down to zero velocity so it does not quite get to the other electrode. In such an experiment, we find the stopping voltage increases linearly as we increase the light frequency (and hence the photon energy).

The explanation of this effect by Einstein requires not only that light is emitted in quanta by objects, which was Planck’s proposal, but that it actually also *consists* of such quanta – the photons of energy $h\nu$.

Quiz questions

- Q7.2.1 If we increase the photon energy in the light shining on the metal plate in the photoelectric effect experiment, we should expect to have to increase the magnitude of the voltage on the collecting plate to stop the flow of electrons. True or False?
- Q7.2.2 If we increase the amount of light power at a specific wavelength that is shining on the metal plate in the photoelectric effect experiment, we should expect to

have to increase the magnitude of the voltage on the collecting plate to stop the flow of electrons. True or False?

Problems

- 7.2.1 Suppose that the work function of some metal is 2eV , and that we illuminate it with blue photons at a wavelength of 425 nm . What minimum stopping voltage should be required to stop the emitted electrons from reaching the collecting plate? Specify also the sign of the stopping voltage, with a positive voltage on the stopping plate relative to the emitting metal being a positive number. [Hints: you do not need to use the mass of the electron nor do you need to know the distance from the emitting metal plate to the collecting plate. How much kinetic energy in electron-volts do you need to “climb uphill” against a stopping voltage of magnitude 1V ?]

7.3 Light and modes 22b

To understand light and quantum mechanics, in addition to the idea of photons, we need to consider modes for light. We introduced before the ideas of modes for standing waves on a string, for example. One approach we take with light is similar to that of waves on a string, but instead we consider waves in a box with perfectly reflecting walls. This also will give us standing waves, but now these are waves that can be “standing” in all three directions.

It might seem somewhat artificial to introduce this box, because we are in reality going to be considering light in situations where there is no such box. This kind of approach is a common fiction in physics, and we believe it is at least useful because it can lead to results that do not depend on the size of the box. Also, we can typically get to equivalent results by presuming traveling waves rather than standing ones, so we rationalize that this use of a box, though artificial, is likely not critical to the underlying behavior. For further simplicity here, we assume the box is cubic, as sketched in Fig. 7.3, with sides of length L . We can also refer to such a box with these perfectly reflecting walls as a “cavity”.

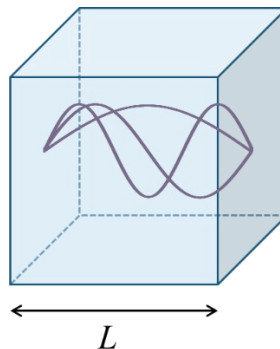


Fig. 7.3. A cubic box or “cavity” of side L , with perfectly reflecting internal walls. Also sketched are sine waves in one direction that fit within the box.

Because we presume perfectly reflecting walls, we argue that we can take the wave to be zero at these walls. Of course, in one direction we can have waves of the form $\sin k_x x$ where k_x would be chosen to fit integer numbers of half-wavelengths within the box, hitting zero at the two walls in that direction. The possible values of k_x would then be

$$k_x = \frac{\pi}{L}, \frac{2\pi}{L}, \frac{3\pi}{L}, \dots, \frac{n_x \pi}{L} \dots \quad (7.1)$$

where n_x is a positive integer.

For three dimensions, we need to generalize, and we propose that the possible modes are of the form $\sin(k_x x) \sin(k_y y) \sin(k_z z)$ in which integer numbers of half waves fit inside the box in each of the three dimensions, so k_y and k_z each have sets of values also like those in Eq. (7.1).

Note that, with expressions like Eq. (7.1) for each of k_x , k_y and k_z , the allowed values for each component k_x , k_y and k_z are spaced by π / L . So, if we are interested in some range of k_x of size Δk_x , we should expect to find

$$\frac{\Delta k_x}{(\pi / L)} = \frac{L}{\pi} \Delta k_x \quad (7.2)$$

different possible k_x values in that range. We can argue similarly for k_y and k_z . So, for ranges Δk_x , Δk_y , and Δk_z , we should expect to find

$$\frac{L}{\pi} \Delta k_x \frac{L}{\pi} \Delta k_y \frac{L}{\pi} \Delta k_z = \frac{L^3}{\pi^3} \Delta k_x \Delta k_y \Delta k_z \equiv g_k \Delta k_x \Delta k_y \Delta k_z \quad (7.3)$$

different possible values of the wave vector

$$\mathbf{k} = k_x \hat{\mathbf{x}} + k_y \hat{\mathbf{y}} + k_z \hat{\mathbf{z}} \quad (7.4)$$

where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are unit vectors in the x , y , and z directions, respectively.

Here we can call

$$g_k \equiv \frac{L^3}{\pi^3} = \frac{V}{\pi^3} \quad (7.5)$$

the density of states in k -space for a volume $V = L^3$.

Below, we will in the end mostly be interested in this density of states per unit volume, so we will be dividing by V . In this case, then, effectively we can regard the size of this box or cavity, and possibly also its shape, as being irrelevant, and we can presume we have effectively derived some more universal density of states per unit volume.¹

¹ For large volumes, this idea of a density of states per unit volume does seem to work, but the argument we have used here (and it is a common one for this problem) is not very satisfactory. It relies on the assertion of some kind of resonator that does not actually exist (the internally reflecting cube). The reason for the use of the resonator is that we can formally solve eigenproblems for resonators, which lead to orthogonal sets of resonator modes, and we need that orthogonality for a meaningful counting. So, since historically we have not had another good

With this example set of modes of a cubic cavity with perfectly reflecting walls, and this idea of the density of states in k -space, we can now start to derive the thermal behavior of light.

Quiz questions

Q7.3.1 Increasing the size of the reflecting cubic box decreases the separation between allowed k values for light-wave cavity modes in the x , y , or z directions. True or False?

Q7.3.2 The density of states in k space for light-wave cavity modes increases in proportion to the volume of the reflecting cubic box. True or False?

7.4 Thermal radiation

In Chapter 5, we derived Planck's distribution, which gives the average number of photons per mode, $\langle q \rangle$ (Eq. (5.98)) in thermal equilibrium at some temperature. Using this $\langle q \rangle$, we can work out the spectrum of that radiation – that is, how much light is there as a function of frequency at a given temperature – and the total amount of energy in the light overall in such a thermal equilibrium distribution. To do so, we need to return to the modes of the light field. We need to put Planck's distribution together with the number and distribution of the modes with frequency.

Planck's radiation law

So, with the average number of photons in a particular mode of angular frequency ω , as given by Planck's distribution, Eq. (5.98) ($\langle q \rangle = 1 / [\exp(\hbar\omega / k_B T) - 1]$), the total energy in the electromagnetic field per unit volume – the energy density U – is given by the sum of the corresponding energies for all the modes

$$U = \frac{1}{V} \sum_{\text{polarizations}} \sum_{k_x} \sum_{k_y} \sum_{k_z} \langle q \rangle \hbar\omega = \frac{2}{V} \sum_{k_x} \sum_{k_y} \sum_{k_z} \langle q \rangle \hbar\omega \quad (7.6)$$

where we have summed over the two polarizations possible for each mode of an electromagnetic field (e.g., horizontal polarization and vertical polarization) to get the factor of 2 here.

We presume the volume or “box” we are considering is large, so the allowed values of the components k_x , k_y , and k_z are very closely spaced. Hence, we can approximate the sum by an integral where we allow each component k_x , k_y , and k_z to range from

way of getting that orthogonality, as physicists we just think (or hope) we are going to get away with this.

There is another way of approaching the problem which does not require assertion of non-existent resonators, and gives orthogonal modes in any volume. Arguably, it gives the correct way out of this problem. See D. A. B. Miller, "Waves, modes, communications, and optics: a tutorial," *Adv. Opt. Photon.* 11, 679-825 (2019) <https://doi.org/10.1364/AOP.11.000679>

approximately zero to infinity, and we use the density of k -states $g_k = V / \pi^3$ of Eq. (7.5). So the total energy per unit volume for all these modes is

$$U = \frac{2}{V} \sum_{k_x} \sum_{k_y} \sum_{k_z} \langle q \rangle \hbar \omega = \frac{2}{V} \int_{k_x=0}^{\infty} \int_{k_y=0}^{\infty} \int_{k_z=0}^{\infty} \langle q \rangle \hbar \omega g_k dk_x dk_y dk_z \quad (7.7)$$

Nothing in the integral in Eq. (7.7) depends on the direction of the wave vector $\mathbf{k}$, so we can change to spherical polar coordinates and then perform our “volume” integral using spherical shells of radius k , surface area $4\pi k^2$, and thickness dk ; we only need one “octant”, that is $1/8$, of that spherical shell because we are only considering positive values of k in each of the three directions. So, in using these spherical shells we need to divide by 8. Hence we have

$$U \simeq \frac{2}{8V} \int_{k=0}^{\infty} \langle q \rangle \hbar \omega g_k 4\pi k^2 dk = \int_{k=0}^{\infty} \langle q \rangle \hbar \omega \frac{1}{\pi^2} k^2 dk \quad (7.8)$$

For light in free space, for angular frequency ω , we have² $k = \omega / c$, so changing variables from k to ω , we have

$$U = \frac{\hbar}{\pi^2 c^3} \int_{\omega=0}^{\infty} \omega^3 \langle q \rangle d\omega \quad (7.9)$$

We can think of this integral, Eq. (7.9), as an integral over a quantity u_ω , that is

$$U = \int_0^{\infty} u_\omega d\omega \quad (7.10)$$

Then u_ω is the energy density per unit (angular) frequency for light in thermal equilibrium at a temperature T . We can write this explicitly (substituting for $\langle q \rangle$ from Eq. (5.98)) as

$$u_\omega = \frac{\hbar}{\pi^2 c^3} \frac{\omega^3}{\exp(\hbar \omega / k_B T) - 1} \quad (7.11)$$

We can call this expression Planck’s radiation law³. This actually gives the form of the “black-body” spectrum, though we have a few more conceptual steps to complete that connection. For the moment, it is still just the energy per unit volume per unit (angular)

² This is equivalent to the relation $f = c / \lambda$ between the frequency f , wavelength λ , and velocity of light c .

³ The form we are giving here is just for the energy density per unit volume. Often, Planck’s radiation law is quoted as a radiated intensity per unit solid angle (or “spectral radiance”). In that case, one divides this expression by 4π to account for the 4π steradians of solid angle, and then also presumes that this energy propagates in the direction(s) of interest at a velocity c (the velocity of light), so we have to multiply by c to get that the resulting intensity and hence spectral radiance, so, overall, multiplying by $c/4\pi$ compared to our formula here.

frequency in a light field in thermal equilibrium at a given temperature. This function is plotted in Fig. 7.4.

Note one important point about this Planck spectrum: the exponential function on the bottom line of Eq. (7.11) grows faster at high frequencies or photon energies than the ω^3 term on the top line, which causes this distribution to fall off at high frequencies. In turn this is because the Boltzmann factor exponentially decreases the probability of the existence of photons of large photon energies, and hence reduces the energy per mode. This is the reason why the quantization of light into photons solves the problem of the “ultraviolet catastrophe” of classical physics, in which the energy in the light would grow without bound as we considered higher and higher frequencies.

We can also consider this energy density in terms of wavelength. Obviously, whether or not we integrate over (angular) frequency ω in Eq. (7.10), or if we change the variable to integrate correctly over wavelength instead, we should get the same total energy density in the electromagnetic field.

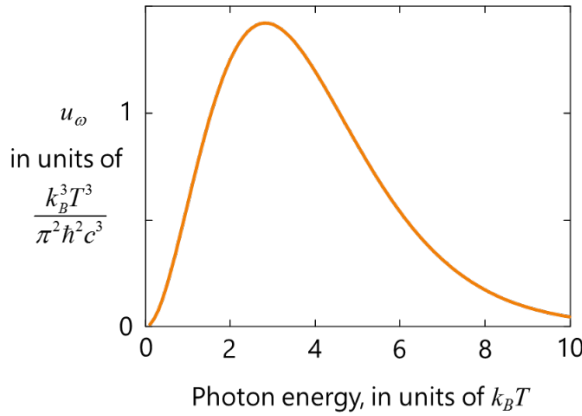


Fig. 7.4. Plot of u_ω , the energy per unit volume per unit angular frequency, in units of $(k_B^3 T^3 / \pi^2 \hbar^2 c^3)$, as a function of the photon energy, in units of $k_B T$.

So, we could define an energy density per unit wavelength, u_λ , so that

$$U = \int_0^\infty u_\lambda d\lambda \quad (7.12)$$

Formally, we can change the variable in the integration in Eq. (7.10) from ω to λ . Noting that

$$\omega = 2\pi f = 2\pi c / \lambda \quad (7.13)$$

we see first that when $\omega=0$, $\lambda=\infty$, and vice versa. So, noting also that $d\omega / d\lambda = -2\pi c / \lambda^2$, we have

$$U = \int_0^\infty u_\omega d\omega = \int_\infty^0 u_\omega \frac{d\omega}{d\lambda} d\lambda = \int_0^\infty \frac{2\pi c}{\lambda^2} u_\omega d\lambda \quad (7.14)$$

So, therefore, we can identify the quantity $u_\lambda \equiv 2\pi c u_\omega / \lambda^2$. All that remains is to rewrite u_ω in terms of wavelength λ rather than angular frequency ω . So, from Eq. (7.11) and substituting from Eq. (7.13), we have

$$u_\lambda = \frac{2\pi c}{\lambda^2} \frac{\hbar}{\pi^2 c^3} \frac{(2\pi c)^3}{\lambda^3} \frac{1}{\exp(\hbar c / \lambda k_B T) - 1} = \frac{8\pi \hbar c}{\lambda^5} \frac{1}{\exp(\hbar c / \lambda k_B T) - 1} \quad (7.15)$$

The Stefan-Boltzmann law

Now we can complete the integration over frequency in Eq. (7.9) (or, equivalently, Eqs. (7.10) and (7.11)) to get the total thermal radiation density

$$U = \int_0^\infty u_\omega d\omega = \int_0^\infty \frac{\hbar}{\pi^2 c^3} \frac{\omega^3}{\exp(\hbar \omega / k_B T) - 1} d\omega = \frac{k_B^4 T^4}{\pi^2 \hbar^3 c^3} \int_0^\infty \frac{x^3}{\exp(x) - 1} dx \quad (7.16)$$

In the last step on the right in Eq. (7.16), we changed variables to $x = \hbar \omega / k_B T$. Using the result

$$\int_0^\infty \frac{x^3}{\exp(x) - 1} dx = \frac{\pi^4}{15} \quad (7.17)$$

we obtain the energy per unit volume inside the cavity

$$U = \frac{\pi^2 k_B^4}{15 \hbar^3 c^3} T^4 \quad (7.18)$$

This result is known as the Stefan-Boltzmann⁴ law. In an electromagnetic field at a given temperature, the energy density is proportional to the fourth power of the temperature.

Quiz questions

- Q7.4.1 The energy density per unit (angular) frequency for light at thermal equilibrium at temperature T increases as the cube of the (angular) frequency. True or False?
- Q7.4.2 The total energy density for light at thermal equilibrium at a temperature T increases as the fourth power of T (i.e., in proportion to T^4). True or False?

Problems

- 7.4.1* Suppose I have a 1 m^3 box, which we can think of as having perfectly reflecting walls. I imagine that this box is near the surface of the sun, so it is at the same temperature as the sun, which we take to be 5800 K. How much energy is stored in the electromagnetic radiation inside the box?

⁴ Josef Stefan (1835–1893)

- 7.4.2 Presume my room has a size $3 \times 4 \times 2.5 \text{ m}^3$, and that the electromagnetic field inside the room is in thermal equilibrium with the walls at a temperature of 25°C (and the room is otherwise completely dark). How much energy, in joules, is there in the electromagnetic field in my room?
- 7.4.3 We can see from Fig. 7.4 that the energy per unit frequency (or angular frequency), Eq. (7.11), in a radiation field at thermal equilibrium at a given temperature is peaking at just below $3k_B T$. In fact, after some work (the calculus does not have a particularly simple form), it is possible to show that this photon energy of peak energy per unit frequency is at $\sim 2.821 k_B T$. Calculate this peak photon energy in electron-volts for the following cases:
- (i) the temperature of the sun ($\sim 5800 \text{ K}$)
 - (ii) a room temperature of $\sim 300 \text{ K}$.
 - (iii) the cosmic microwave background radiation from the universe ($\sim 2.7 \text{ K}$)

7.5 Black-body radiation

So far, we have shown how to calculate the spectrum and total energy in an electromagnetic field at thermal equilibrium at a given temperature. While interesting to some degree, these results may not appear particularly useful. Once we can relate these results to the limits on the absorption and emission of light from hot bodies, they become quite directly useful, however, allowing us to understand limits to light bulbs, for example, and the spectrum and power of the light from the sun. Both of these topics are quite important for energy consumption and generation more generally.

To make the connection between the Planck radiation law and the Stefan-Boltzmann law on the one hand, and the light emitted and absorbed by bodies on the other, we need to introduce the idea of a “black body”. As we will see, this will be both a perfect absorber and a perfect emitter of light.

Suppose now that we make a small hole in the cavity, as shown in Fig. 7.5(a), so some of the radiation can escape. We can make this hole as small as we wish, to make this perturbation of the system as small as we want. So, we believe the spectrum of the light emitted by this hole will be the same as that of the electromagnetic field in thermal equilibrium at the same temperature as the cavity.

Next, suppose we shine light into this cavity through this small hole, as shown in Fig. 7.5(b). Because the hole is small, the light will bounce around many times inside the cavity. Assuming at least some small absorption inside the cavity, this light will be almost entirely absorbed and not re-emitted through the hole. Such an object is known as a “black-body”, because, at least in this hole, it is a perfect absorber of light.

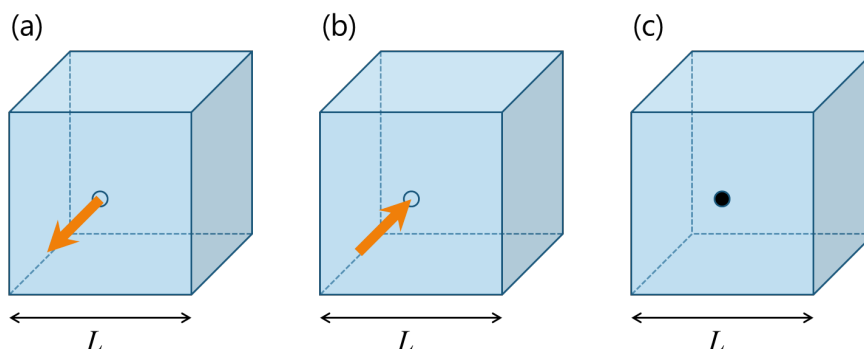


Fig. 7.5. A cavity with a small hole, (a) considering light emitted through the hole, (b) considering light shining into the hole (to be absorbed), and (c) covering the hole with a black (perfectly absorbing) surface.

Because this is a cavity at thermal equilibrium, the light inside has the spectrum given by Planck's radiation law, so the light emitted through the hole, called "black body radiation", has this spectrum also.

Suppose now we close off this hole with a "black" (perfectly absorbing) surface, as shown in Fig. 7.5(c). In thermal equilibrium, the radiation emitted by this black surface into the resonator must equal the radiation landing on the black surface from the inside of the resonator. Hence the thermal radiation emitted from a black surface is equal to the thermal radiation from a small hole in a cavity (at the same temperature). Hence the Stefan-Boltzmann and Planck radiation laws can allow us to calculate the thermal radiation emitted from a black surface.

One key point about "black bodies" is that they set a limit on the emission of light by hot bodies in thermal equilibrium. Suppose that we presumed there was some body that at thermal equilibrium could emit more electromagnetic radiation than a black body. If there were such a body, we could heat up the cavity, which starts at the same temperature as the body, solely with the radiated thermal energy from that body. Presuming we could do this to some useful degree, then the cavity would become hotter and so we would have heat flowing from a cooler body to a hotter body in a closed system, which would violate the second law of thermodynamics.

These results from the thermal physics of light are fundamentally bad news for incandescent light bulbs. There is a limit to how much light we can get out of something just by heating it up, and there is nothing we can do about this limit. If we want more light from a given size of incandescent bulb, we have to make it hotter, and materials can only be made so hot before they evaporate.

Kirchhoff's law of radiation

So far, we have only considered the necessary thermal radiation from a perfectly absorbing "black" body. We can, however, also come to some useful conclusions about "non-black" objects, ones that emit less thermal radiation about some given wavelength than a black body.

For such a body, we can call the ratio of its emitted energy to that of a black body the “emissivity”, e_s . This emissivity can be a function of wavelength. Because we know that no body can have greater thermal emission at any wavelength than a black body, we know that $e_s \leq 1$. Suppose now that this object absorbs a fraction a of the incident radiation at or about the wavelength of interest. We can call this fraction a the “absorptivity”.

But, in thermal equilibrium, the body must emit the same amount of radiation that it absorbs. To see this, imagine we set up this body so that it can exchange radiation with a black body at the same temperature. If the body absorbed more than it emitted, then the black body at the same temperature could heat it up. If the body emitted more than it absorbed, then it could heat up the black body. Both of these supposed processes violate the second law of thermodynamics by heading the closed system towards a state of lower entropy, with resulting different temperatures in the two bodies.

Hence we conclude that

the absorptivity and the emissivity are equal

which is Kirchhoff’s law of radiation⁵.

Quiz questions

- Q7.5.1 Suppose I want to stay warm at night, and I have a choice of two blankets that are the same in all respects except that one is white and the other is black. Which blanket should I choose?
- Q7.5.2 Suppose I want to warm up some object from some cold temperature using sunlight. Should I paint it white or black?
- Q7.5.3 Suppose I want my house to cool down at night. Should I paint my roof white or black?
- Q7.5.4 Suppose I have two objects, one white and the other black that are initially at the same temperature. For the purposes of this question, we assume there is no other source of heat other than electromagnetic radiation from one or other body, and there is no other “sink” of heat other than absorption or electromagnetic radiation by one or other body. (There is, therefore, no other conduction of heat between the two bodies.) I put them reasonably close to one another. Choose one of the following outcomes:
- a) The black one will get warmer because it absorbs light emitted by the white one.
 - b) The white one will get warmer because it absorbs light emitted by the black one.
 - c) Neither one will change temperature.

⁵ For a recent discussion and extension of the ideas of this law, see also D. A. B. Miller, L. Zhu, and S. Fan, “Universal modal radiation laws for all thermal emitters,” PNAS 114, 4336-4341 (2017) <https://doi.org/10.1073/pnas.1701606114>

7.6 Absorption and emission of light

In modern quantum mechanics, we have very good microscopic models of how light interacts with matter. The detail of those models goes beyond what we can present here; to complete them requires techniques such as perturbation theory and a more advanced quantum theory of light itself, though all of this is relatively straightforward even if it would take some work.

In 1916-17, however, long before we had any such detailed quantum mechanical models of light itself and how it interacts with matter, Einstein made some simple assumptions and derived a remarkable result that anticipates much of the subsequent theory of such processes. This is known as Einstein's A & B coefficient argument⁶. It is remarkable not only for showing how absorption and emission are related in quantum mechanics, but also in predicting the process of stimulated emission. That process, which would not be exploited for about another 50 years, lies behind the whole concept of lasers, which we use widely today for many applications.

Here, we start with Planck's radiation law, Eq. (7.11), which we wrote as the energy per unit volume per unit (angular) frequency. For this argument, we will use the version of that law using ordinary frequency ν rather than angular frequency, so we substitute using $\omega = 2\pi\nu$ in Eq. (7.11), and since we want energy per unit frequency rather than per unit angular frequency, we also multiply the previous formula, Eq. (7.11), by 2π .

Hence we have the energy per unit volume per unit frequency in the electromagnetic field at thermal equilibrium, which we write as

$$\rho(\nu) = \frac{8\pi h\nu^3}{c^3} \frac{1}{\exp(h\nu / k_B T) - 1} \quad (7.19)$$

Now, to construct this argument, we presume we have some atoms, all the same kind, and that they have two possible states, state 1 of energy E_1 and state 2 of (greater) energy E_2 , as sketched in Fig. 7.6(a).

In this collection of atoms, we have N_1 (per unit volume) in state 1 and N_2 (per unit volume) in state 2, with everything – atoms and the electromagnetic radiation – in thermal equilibrium at temperature T . We also expect in thermal equilibrium that the ratio between N_2 and N_1 should be given by the Boltzmann factor, so that

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{k_B T}\right) \quad (7.20)$$

We presume the only way atoms change states from state 2 to state 1 is by emitting a photon of energy $h\nu$ where

$$h\nu = E_2 - E_1 \quad (7.21)$$

⁶ For a recent extension of Einstein's A & B coefficient argument to a modal version, see D. A. B. Miller, "Waves, modes, communications, and optics: a tutorial," *Adv. Opt. Photon.* 11, 679-825 (2019) doi:10.1364/AOP.11.000679

and similarly the only way the atoms change states from state 1 to state 2 is by absorbing a photon of energy $h\nu$.

Because of Eq. (7.21), we can rewrite the Boltzmann factor as

$$\frac{N_2}{N_1} = \exp\left(-\frac{h\nu}{k_B T}\right) \quad (7.22)$$

We expect there is an emission process – spontaneous emission – by which an atom in state 2 will emit a photon and fall to state 1. We expect this process happens independent of the amount of light also present. We propose the number of such spontaneous emission transitions is simply proportional to the number of atoms in state N_2 , and we propose a constant coefficient A_{21} – an intrinsic property of the atom, not dependent on temperature – giving a spontaneous emission rate $A_{21}N_2$, as sketched in Fig. 7.6(b).

We propose also that there is an absorption process resulting from absorbing photons from the electromagnetic field, and we propose an absorption rate from the N_1 atoms in state 1 for taking them to state 2 that is proportional to amount of electromagnetic radiation per unit frequency of frequency near to ν – in other words, this absorption rate should be proportional to $\rho(\nu)$ in thermal equilibrium – and we presume some constant coefficient B_{12} for this process that is an intrinsic temperature-independent property of the atoms. So we have a total absorption rate $B_{12}N_1\rho(\nu)$, as sketched in Fig. 7.6(c).

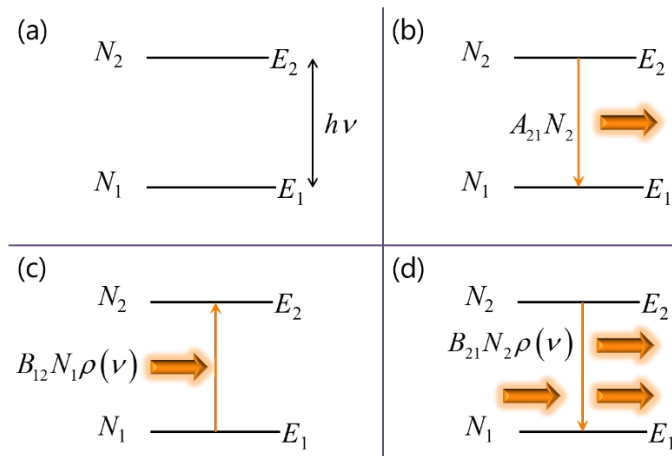


Fig. 7.6. (a) Illustration of atoms with two levels, separated by energy $h\nu$. (b) The process of spontaneous emission of photons. (c) The process of absorption of photons. (d) The process of stimulated emission of photons.

These proposals so far are perfectly reasonable. If atoms are in their excited state, then it is reasonable to say they will fall to their lower state, emitting light, even in the dark – we see that all the time with hot bodies – and it is quite reasonable to say that this (spontaneous) emission process still happens in the same way even if we shine a light on the material. We might change the populations of the states by shining the light on the system, but still we can reasonably expect that the strength of this process itself is

not changed by doing so. Similarly, there is some absorption process for light that has a strength determined by intrinsic properties of the atoms.

Now, a basic presumption of statistical mechanics is that we have “detailed balance”: in equilibrium, we must have exactly equal rates for processes in both directions. In equilibrium, if there is some process that excites systems, then there must be another process that de-excites them, and the rates must balance overall, otherwise we would not be in equilibrium.

But this detailed balance gives a problem if we stop our model here. Asking for detailed balance so far means we would be equating the spontaneous emission and the absorption for our thermal equilibrium system of atoms and radiation, which would give

$$A_{21}N_2 = B_{12}N_1\rho(\nu) \quad (7.23)$$

Using Eqs. (7.19) and (7.22), we obtain

$$\frac{A_{21}}{B_{12}} = \frac{N_1}{N_2} \rho(\nu) = \frac{8\pi h \nu^3}{c^3} \frac{\exp(h\nu/k_B T)}{\exp(h\nu/k_B T) - 1} = \frac{8\pi h \nu^3}{c^3} \frac{1}{1 - \exp(-h\nu/k_B T)} \quad (7.24)$$

By presumption, A_{21} and B_{12} are intrinsic coefficients for the atoms, and they do not depend on temperature, but the right hand side of this equation does depend on temperature, so there is an apparent contradiction.

Einstein realized a way out of this difficulty. He proposed another process – stimulated emission – a process just like absorption, but from state 2 to state 1. He proposed another coefficient B_{21} , also presumed to be an intrinsic temperature-independent atomic property. Hence we obtain an additional emission rate $B_{21}N_2\rho(\nu)$, as sketched in Fig. 7.6(d). Now detailed balance gives

$$A_{21}N_2 + B_{21}N_2\rho(\nu) = B_{12}N_1\rho(\nu) \quad (7.25)$$

where the emission terms are on the left, and the absorption terms are on the right. Rewriting Eq. (7.25) gives

$$A_{21} = \left(B_{12} \frac{N_1}{N_2} - B_{21} \right) \rho(\nu) = \frac{8\pi h \nu^3}{c^3} \frac{[B_{12} \exp(h\nu/k_B T) - B_{21}]}{\exp(h\nu/k_B T) - 1} \quad (7.26)$$

So if we choose

$$B_{21} = B_{12} \quad (7.27)$$

then we can have detailed balance independent of temperature, and we also obtain the relation

$$A_{21} = \frac{8\pi h \nu^3}{c^3} B_{12} \quad (7.28)$$

So, in summary, if we presume intrinsic microscopic processes for spontaneous emission, with rate $A_{21}N_2$, and absorption, with rate $B_{12}N_1\rho(\nu)$, then to achieve detailed balance we need to propose stimulated emission, with rate $B_{21}N_2\rho(\nu)$, and we find the results Eqs. (7.27) and (7.28).

Since Eq. (7.28) tells us that the coefficient A_{21} for spontaneous emission just involves fundamental constants, the frequency ν , and the coefficient B_{12} , the same microscopic atomic coefficient underlies all three processes of absorption, spontaneous emission, and stimulated emission. This is a remarkable result that ties together all the absorption and emission processes.

We can derive one other simple result here, which is the ratio of spontaneous and stimulated emission. We note we can write the energy per unit volume per unit frequency, as in Eq. (7.19), as

$$\rho(\nu) = \frac{8\pi h \nu^3}{c^3} n_{ph} \quad (7.29)$$

where n_{ph} , which we now think of as the number of photons per mode for a mode of frequency ν , is just the same as the quantity $\langle q \rangle$ from the Planck distribution, Eq. (5.98) ($\langle q \rangle = 1 / [\exp(\hbar\omega / k_B T) - 1]$). Explicitly now, using frequency rather than angular frequency, Eq. (5.98) becomes

$$n_{ph} = \frac{1}{\exp(h\nu / k_B T) - 1} \quad (7.30)$$

The total spontaneous emission rate per unit volume per unit frequency is, therefore, using Eq. (7.28)

$$R_{sp} = A_{21} N_2 = \frac{8\pi h \nu^3}{c^3} B_{12} N_2 \quad (7.31)$$

and the total stimulated emission rate per unit volume per unit frequency is

$$R_{stim} = B_{12} N_2 \rho(\nu) = \frac{8\pi h \nu^3}{c^3} B_{12} N_2 n_{ph} \quad (7.32)$$

So

$$\frac{R_{stim}}{R_{sp}} = \frac{(8\pi h \nu^3 / c^3) B_{12} N_2}{(8\pi h \nu^3 / c^3) B_{12} N_2} n_{ph} = n_{ph} \quad (7.33)$$

where we have also used Eq. (7.27). So we come to the simple conclusion that the ratio of stimulated to spontaneous emission in a given mode is the number of photons in the mode.

Under normal circumstances, as in the light from the sun, there is not much stimulated emission present, though it is not quite negligible. For visible radiation, such as for a photon energy ~ 2.5 eV, corresponding to a wavelength of ~ 500 nm, the number of photons per mode at 5800 K (the approximate temperature of the sun) directly from the sun is

$$\begin{aligned}\frac{1}{\exp(h\nu/k_B T) - 1} &= \frac{1}{\exp(2.5e/k_B T) - 1} \approx \exp\left(-2.5\frac{e}{k_B T}\right) \\ &\approx \exp\left(-2.5\frac{1.602 \times 10^{-19}}{1.38 \times 10^{-23} \times 5800}\right) \approx \exp(-5) = 0.007\end{aligned}\quad (7.34)$$

so about 0.7% of the sun's photons at 500 nm are from stimulated emission.

For light from the sun, we might not regard such an amount of stimulated emission as being very important, even if it is necessary for the statistical mechanics of light to work out correctly. A key point, however, is that, though we derived Eq. (7.33) so as to make detailed balance work at thermal equilibrium, we expect this process of stimulated emission and the result Eq. (7.33) to remain valid even when we are not at thermal equilibrium. In the case of a laser, we will have very large numbers of photons in the “lasing” mode, which means the laser emits almost entirely through stimulated emission, putting yet more photons into the same mode, and giving us laser beams.

Quiz question

- Q7.6.1 Suppose we have a gas of atoms, each of which has two possible energy states, E_1 and E_2 , with $E_2 > E_1$. Suppose N_1 of the atoms are in the state with energy E_1 and N_2 are in the state with energy E_2 . Suppose specifically for the moment that $N_1 = N_2$.
- Suppose there are no photons present. State whether the following statement is True or False: Then there will be no emission or absorption of light.
 - Suppose now there are photons present in a light beam, at least some of which have an energy $\hbar\omega$ corresponding to the energy separation $E_2 - E_1$. State whether the following statement is True or False: Then there will be no net absorption of photons from the light beam.
 - Suppose now that $N_2 > N_1$. State whether the following statement is True or False: It is possible that if I come in with a light beam, the outgoing light beam could have more photons in it than the incoming light beam.

Problem

- 7.6.1 Suppose we have a gas of atoms, each of which has two possible energy states, E_1 and E_2 , with $E_2 > E_1$. Suppose that all the atoms are in their upper (E_2) state. I come into this gas with n photons, all of an appropriate energy to cause transitions between these levels. All these n photons are in one specific mode. Then, the total rate of emission of photons by the gas into this mode is (on the average) which one of the following possibilities (justify your answer briefly)?
- independent of n
 - proportional to n

(c) proportional to $n + 1$

[Hint: Look very closely at Eqs. (7.31) to (7.33), and remember we are looking for the total emission rate into this mode.]

7.7 Conclusions

In this chapter, we have seen that, with the basic presumption that light exists as photons of energy $E = h\nu \equiv \hbar\omega$, we can put this together with our general understanding of thermal distributions to construct a powerful model of light. We can explain the “black-body” spectrum that was such a mystery at the end of the 19th century, and we can generally explain radiation by hot bodies and the basic relation between emission and absorption. With Einstein’s A and B coefficient argument, we can even explain the process of stimulated emission, the core process behind all lasers. With these principles, we are now in a position to understand a wide range of optical and optoelectronic devices and applications.

Chapter 8

Semiconductor optoelectronics

8.1 Introduction 24a

Our ability to work with light has been transformed by devices that enable us quickly and efficiently to convert between electrical and optical energy and signals. The devices that enable such conversions are called optoelectronic devices. Viewed broadly, that category includes conventional incandescent and fluorescent light bulbs, vacuum-based devices like cathode ray tubes and plasma displays and some specialized photodetectors¹, and liquid crystal display technology. A large fraction of all modern optoelectronic devices are, however, based on semiconductors, and these include a broad range of photodetectors (including solar cells), optical modulators, light-emitting diodes, and lasers. Here, we will introduce the physics behind such semiconductor optoelectronics, building on the principles from previous chapters.

Many of these devices are built using diodes. In electronics, diodes are used first for their rectifying properties. In semiconductor optoelectronics, however, diodes are used for quite different reasons. First, they can collect the current of electrons and holes that are generated by absorbing photons inside the diodes; we use this both for solar cells and for detecting signals on light beams. Second, forward biasing the diodes injects electrons and holes into the junction region of the diodes, and those electrons and holes can recombine by emitting photons; we use this process in the light-emitting diodes that give us efficient sources of illumination, and in semiconductor lasers that we use both for generating the light for powering optical fiber communications and as an efficient primary source for high power lasers generally. Third, reverse biasing diodes provides an efficient way of applying large electric fields inside devices without much current; we use this for several kinds of light modulators to turn light beams on and off in optical communications – these large electric fields can change the material's optical

¹ Photomultiplier tubes and related devices like image intensifiers can be in this vacuum photodetector category.

properties, such as its optical absorption strength or its refractive index. Here we will introduce especially the use of diode structures in detection and light emission.

8.2 Photodetectors and solar cells

Photoconductors

In a semiconductor, absorbing a photon whose photon energy exceeds the band gap energy can take an electron from the valence band and put it in the conduction band, leaving a hole in the valence band. We can call this process generation of an “electron-hole pair”. The electrons and holes generated by such absorption of a photon “across” the band gap energy are sometimes called “photocarriers” and we refer to them as being “photogenerated”.

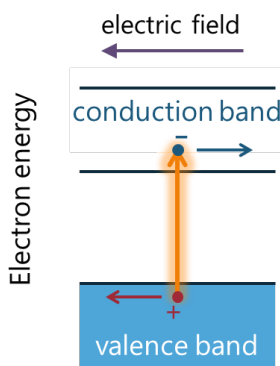


Fig. 8.1. Absorbing a photon of energy greater than the band gap energy can create an electron in the conduction band and a hole in the valence band, which can conduct electricity in the presence of an electric field until they recombine.

It can take some time² for the electron and the hole to “recombine”, with the electron falling back into the valence band. Before they recombine, they can conduct electricity if we apply an electric field, such as by applying a voltage between one end of the semiconductor and the other. In a piece of semiconductor, this leads to what is called photoconduction, as illustrated in Fig. 8.1. This process is used for some photodetectors.

Photodiodes

If we absorb a photon to generate an electron-hole pair inside the depletion region of a diode, as shown in Fig. 8.2, the electron will drift or diffuse “downhill” into the n -doped

² Recombination lifetimes for electrons and holes in semiconductors can be as long as milliseconds or as short as picoseconds. There are many mechanisms for such recombination, including radiative recombination in which a photon is emitted in the recombination process, and recombination through “trap” states (often called “non-radiative” recombination).

region, and the hole will similarly drift or diffuse “downhill” (in its terms³) into the p -doped region.

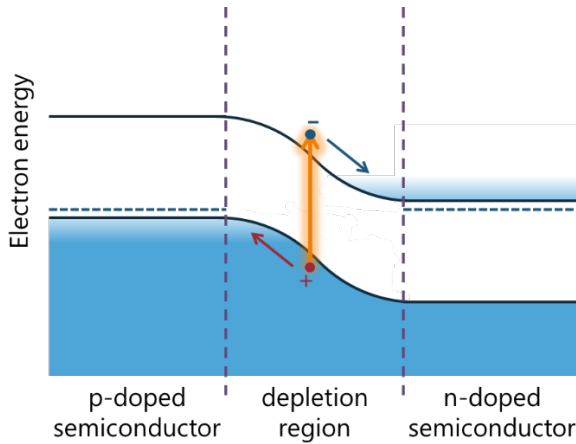


Fig. 8.2. Absorbing a photon across the band gap inside the depletion region leads to the photogenerated electron and the hole moving towards the n - and p -doped regions, respectively.

The charge separation from the electron and hole moving in opposite directions creates an electrostatic potential (or equivalently an electric field) in such a direction as to resist the charge separation; positive charges build up in the p -doped semiconductor and negative charges build up in the n -doped semiconductor, each of which acts to repel the movement of the corresponding photocarrier, at least partly.

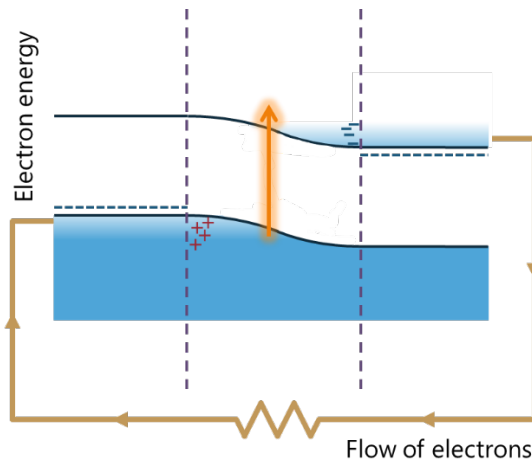


Fig. 8.3. Shining light into the depletion region of a photodiode generates a voltage that can drive current through an external resistor.

³ “Downhill” for holes is “uphill” on the electron energy diagram. We can view hole energies by looking at these electron energy diagrams “upside down”.

This charge separation and the resulting electric potential therefore create an external voltage, which can drive a current through a resistor, as sketched in Fig. 8.3. When we generate voltage from a diode like this, we call this the “photovoltaic” operation of a diode. This process is the basic mechanism in a semiconductor solar cell, generating power by absorbing photons to generate both voltage and current. Such photovoltaic diodes can also be used as photodetectors, generating voltage and current in response to some input light signal.

It is also common to use reverse biased photodiodes as photodetectors, as sketched in Fig. 8.4. Such reverse-biased devices do not generate power in the same way as the photovoltaic diodes do. With reverse bias, however, the electrons and holes are swept particularly quickly out of the depletion region by the large electric field in there, giving high-speed devices. This rapid sweep-out also means that we get nearly exactly an electron of current round the circuit for every photon absorbed, with very few recombining with holes inside the depletion region before being swept out.

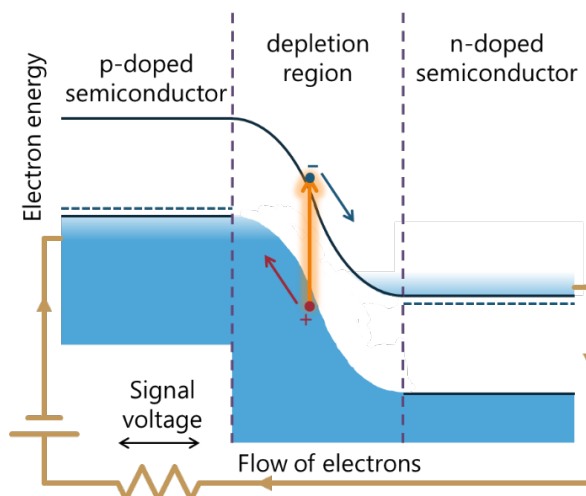


Fig. 8.4. A reverse-biased diode efficiently and rapidly sweeps photocarriers out of the depletion region, allowing a signal voltage across the resistor that is proportional to the incident light.

One other benefit of the reverse-biased operation is that the signal voltage across the resistor can be proportional to the amount of light (number of incident photons per second) shining on the diode. In the photovoltaic case, by contrast, the output voltage tends to be approximately the logarithm of the incident light power because of the underlying exponential behavior, and hence the output voltage is not proportional to the input optical power.

We can now understand the current vs. voltage characteristic of a photodiode, as sketched in Fig. 8.5. In the simplest model, the “photocurrent” – the current from the movement of the photogenerated electrons and holes – behaves like a reverse current; photogenerated electrons and holes are moving “downhill” into their respective *n*- and *p*-doped regions, which is the direction opposite to the forward diffusion current in a diode.

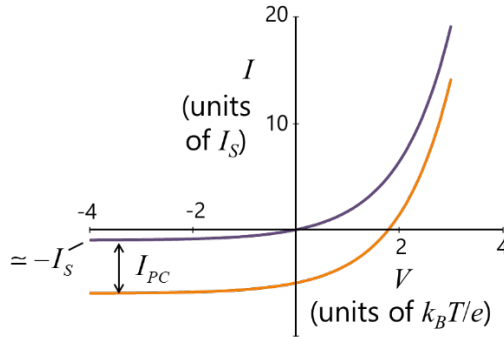


Fig. 8.5. Current vs. voltage curves for a photodiode in the “dark” without incident light (upper curve) and with incident light (lower curve).

So, presuming the photogenerated electrons and holes do get to their respective n - and p -doped regions without recombining, we expect simply to subtract the magnitude of this “photocurrent” I_{PC} from the “dark” diffusion current, shifting the curve down by the photocurrent I_{PC} , and giving a net current

$$I = I_s \left[\exp\left(\frac{eV}{k_B T}\right) - 1 \right] - I_{PC} \quad (8.1)$$

Note that a solar cell would operate in the lower right quadrant of the current vs. voltage characteristic of Fig. 8.5. This is the quadrant in which the diode would generate a power of magnitude $|V \times I|$ in an external resistor in a circuit like that in Fig. 8.3.

Quiz questions

- Q8.2.1 After we excite a semiconductor photoconductor, the additional electrical conductivity will continue until the photocarriers have recombined. True or False?
- Q8.2.2 If I shine light on a photodiode, I will only get voltage across the “ends” (i.e., p and n regions away from the junction) of the photodiode if I connect the two ends through an electrical circuit. True or False?
- Q8.2.3 Current generated in photodiodes as a result of shining light on them is in the same direction as the forward current in a diode. True or False?
- Q8.2.4 In a reverse-biased photodiode, the photocurrent results entirely from diffusion of electrons into the n -doped semiconductor and diffusion of holes into the p -doped semiconductor. True or False?
- Q8.2.5 In an ideal diode operating as a photodiode with light shining on it and being absorbed to generate electrons and holes to create a photocurrent I_{PC} , there will be some light power such that the photocurrent cancels the reverse leakage current of the diode so there can be exactly no reverse current. True or False?

- Q8.2.6 When a photodiode is generating power, the voltage on the diode is in the same sense (i.e., direction) as a forward bias voltage, and the current is in the same sense (i.e., direction) as the reverse leakage current. True or False?

Problems

- 8.2.1* At equilibrium, with zero applied bias voltage in a semiconductor diode, and in the absence of light, the electron flux from the p region to the n region is the same as the electron flux from the n region to the p region, and similarly for holes. Thus, the total net electrical current is zero under such conditions.
- Presume that this diode is at zero bias (i.e., $V_A = 0$). Now we shine light, with photon energy greater than the band-gap energy in the material used to make this diode, into the depletion region, where some of it will be absorbed, creating (photogenerated) electrons and holes. For the electrons and holes created in the depletion region, draw the movements of electrons and holes on the band diagram. Is the photogenerated electrical current positive or negative? (Note: consider positive current to be current in the same direction as that in a forward-biased diode.) What is the driving force for any movements of photogenerated electrons and holes in the depletion region?
 - Based on your answer from part (a), should we apply a positive or a negative voltage to obtain zero current? (Note: positive voltage corresponds to the sign of voltage that would forward-bias the diode.) Explain your answer. (Note: the voltage applied to reach zero current is called the “open-circuit” voltage.)
- 8.2.2 Suppose that we have an “ideal” diode with a saturation current $I_S = 100 \text{ nA}$. We presume that we can shine light into it in such a way that any light that is absorbed is absorbed in the depletion region of the diode. We presume that the semiconductor material of the diode has a bandgap energy of $E_g = 1 \text{ eV}$ (i.e., the minimum separation between the valence and conduction bands is 1 eV), and that any photons of energy $h\nu \geq E_g$ are absorbed. [Any photons of energy $h\nu < E_g$ are not absorbed, though we will not use this fact in this question.] Each photon that is absorbed is presumed to give rise to the equivalent of one electron of “photocurrent” in the circuit. We are working at a “room” temperature of 293 K .
- [Note: for this problem, you may find it useful to use some computer program to graph the appropriate expressions for you. The optimum conditions requested can be deduced just from looking at the graphs, if necessary on some “zoomed-in” portion of them; algebraic proofs of those conditions are not required.]
- Suppose that we are shining sufficient light on this diode that we are generating a photocurrent of 10 mA . [Note: this is not necessarily the total current in the diode – there will also be the usual diffusion current in the diode – though this photocurrent would be close to the total current in the diode if we were running the diode at a large reverse bias.] We presume that we have connected the two

ends of the diode (i.e., the “outsides” of the p and n regions) together using only a resistor. We will later deduce what value, R , we want to choose for this resistor.

Our ultimate goal in this question is to find the conditions of operation of this photodiode so that it is most efficient as a “solar cell” in converting the incident light to electrical power, which we will then dissipate in the resistor. Until we make the final choice of R , we can presume experimentally that we can adjust its value as needed.

- a) Graph the current-voltage characteristic of this diode when there is this much photocurrent (10 mA). [Graphing from 0 to 0.28 V should be a suitable range.]
- b) Graph the power dissipated in the load resistor as a function of the voltage that would be across it. [Again, a range from 0 to 0.28 V should be suitable.] [Note, incidentally, that this power should be positive. The power in the diode – a product of the current through it and the voltage across it – would give a negative number, because it is generating power, not dissipating it. The power dissipated in the resistor should be -1 times this.]
- c) By graphing over an appropriate smaller range (or by some other method if you prefer), find (to at least two significant figures)
 - (i) the maximum power, $P_{\max}$, that could be dissipated in the resistor
 - (ii) the voltage, $V_{\max}$, across the diode when the maximum power is being dissipated in the resistor.
- d) What would be the value of the resistor R to achieve this maximum power, $P_{\max}$? (Two significant figures is sufficient here.) [Note: you can deduce this from the voltage $V_{\max}$ and a knowledge of the dissipated power and/or the corresponding total current at that voltage.]
- e) Supposing that this photocurrent is being generated by a light beam with photons of photon energy 1 eV (all of which are being absorbed in the depletion region of the diode).
 - (i) What is the power in the light beam?
 - (ii) What is the efficiency of this solar cell in converting light power to electrical power in the resistor (when operating at the maximum efficiency conditions derived above)?
- f) Suppose instead the light beam has photon energy 1.5 eV, though it is still generating the same photocurrent.
 - (i) What is the power in this light beam?
 - (ii) What now is the efficiency of converting light power to electrical power in the resistor?
- g) If you could redesign the diode in such a way as to reduce $I_s = 100$ nA to $I_s = 10$ nA, would the efficiency of the diode in converting light to electrical power increase or decrease? [Note: you can prove and/or justify this just by appropriate calculations – algebraic proofs are not required.]

8.3 Light-emitting diodes and lasers

Thermal and non-thermal light sources

We know that heating objects can cause them to emit light as, for example, in an incandescent light bulb. Analyzing the behavior of light emission and absorption in such hot bodies at thermal equilibrium allowed us also to deduce relations between absorption and emission processes through Einstein's A and B coefficient argument. That argument tells us relations between the microscopic processes of absorption and emission in quantum systems like atoms or even states in semiconductor bands, based only on relatively simple assumptions. Those relations between these microscopic quantities remain valid even if we are not at thermal equilibrium, and therefore suggest to us that we can make good light emitters without necessarily making them hot; the key is somehow to get the upper levels of quantum systems populated by some means other than simply heating the systems up.

Until we understood and could control the excitation of quantum mechanical systems by means such as applying voltages, we had no real alternatives just to heating objects up to get them appropriately excited. One very useful technique to achieve such excitation electrically is to use electrical discharges in gasses, which can lead to the gas molecules or atoms being in excited states without necessarily being very hot; this approach enables fluorescent light bulbs, for example, which run at relatively cool temperatures and are significantly more efficient than incandescent light bulbs.

As our understanding of semiconductors improved and we became capable of making diode structures in a variety of materials, we deduced how to make diodes emit light when we forward bias them. This approach has become the basis for light-emitting diodes for illumination and many other applications, for semiconductor lasers for communications, and for efficient and convenient laser sources generally.

Direct band gaps and light emission

We have understood for some time the concept of electron and hole injection into the junction region of diodes. We will discuss that process below, and it does indeed provide an efficient, and non-thermal, method of populating higher-energy quantum mechanical states. But, the advent of efficient diode-based light emitters had to await the ability to make diodes in semiconductors other than the germanium and silicon that were used for early electronics.

In part because they are single-element materials, germanium and silicon can be somewhat easier to make as crystals than other multiple-element materials such as the III-V semiconductors. For a variety of reasons, silicon remains the basis of most semiconductor electronics today, and the associated technology for working with silicon is very advanced. Both germanium and silicon have a problem, however, for light emission: they are both indirect-gap semiconductors.

When we analyze the microscopic processes of light absorption and emission in crystalline materials, we find that the strong processes involve electrons and holes in

the same (or very nearly the same) $\mathbf{k}$ states in different bands. That is, the absorption process that would take an electron in a given $\mathbf{k}$ state in the valence band and put it into essentially the same $\mathbf{k}$ state in the conduction band (what we can call a “vertical” transition) is a particularly strong one; from Einstein’s A and B coefficient argument, we understand also that therefore the emission process that would take an electron in a given $\mathbf{k}$ state in the conduction band and emit a photon to “fill in” a hole at essentially the same $\mathbf{k}$ state in the valence band is a similarly strong process.

One common rationalization of this so-called “direct” absorption process is to argue that the photon has a very small momentum compared to that of the scale of crystal momenta when we look at states in a Brillouin zone, and so to “conserve momentum”, optical transitions between bands have to be essentially “vertical” on such a diagram. If we perform a more detailed analysis of such absorption processes, we do indeed find that this effective momentum can be viewed as being conserved in such a process, though that is a conclusion that comes out of the full analysis, not a principle we put into it.

When we are considering absorption, such “direct” or “vertical” transitions between the valence and conduction band states are quite possible also in indirect band gap materials like silicon or germanium. In absorption, we start with essentially full valence bands and essentially empty conduction bands, so for each full valence band state there is an empty conduction band state “above” it, and so such “direct” absorption transitions are quite possible. Such indirect materials can be good absorbers of light, and silicon is very widely used as a photodetector.

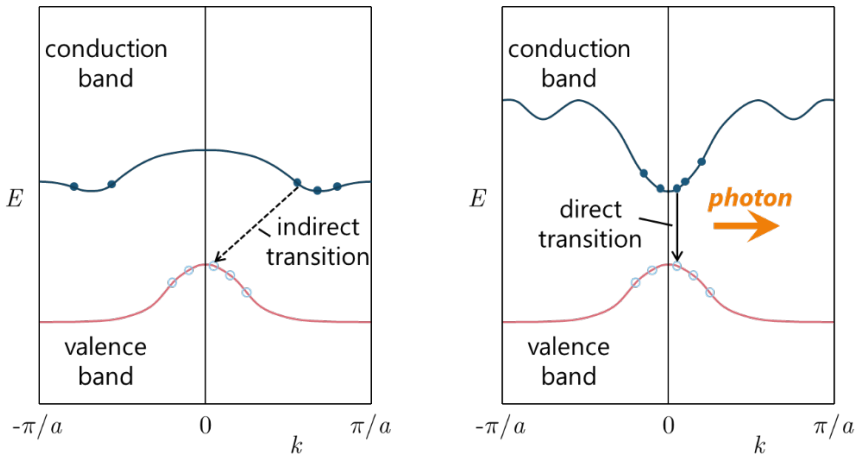


Fig. 8.6. An indirect band gap material (left) showing weak indirect transitions, and a direct band gap material (right) showing strong direct transitions that can emit photons efficiently, both shown in a simple “one-dimensional” Brillouin zone.

When, however, we inject electrons into the conduction band and holes into the valence band, in silicon and germanium, holes collect in one part of the Brillouin zone and electrons collect in other parts, as illustrated on the left of Fig. 8.6; the holes collect in the valence band maxima that are at the center of the zone (at $k = 0$), but the electrons collect in minima that are at or near the edge of the Brillouin zone in these particular

“indirect gap” materials. So where there are electrons in states in the conduction band, there are no holes in the corresponding $\mathbf{k}$ states in the valence band, and where there are holes in states in the valence band, there are no electrons in the corresponding states $\mathbf{k}$ in the conduction band.

III-V materials like GaAs and InP and many others have “direct” band gaps in which the minimum in the conduction band lines up above the maximum in the valence band. Hence when we inject electrons into the conduction band and holes into the valence band, they are in the same regions of the Brillouin zone, as shown in the right of Fig. 8.6. Now if we find an electron in a given $\mathbf{k}$ state in the conduction band, we may well find a hole in the corresponding $\mathbf{k}$ state in the valence band, and so we have strong possibilities for emission of a photon as the electron “falls” down into that $\mathbf{k}$ state in the valence band.

It is possible to make “indirect” transitions that emit a photon and that do not conserve crystal momentum for the electrons, but these are weak⁴. The electron may be more likely to recombine with a hole through some other, “non-radiative” process before these weak radiative processes have much of a chance to happen, so such “indirect” light emission is generally an inefficient process.

Forward-biased diodes for light emitters

As we have already understood from previous discussions of diodes, forward biasing them “injects” additional electrons from the n -doped region on the right into the junction region and similarly injects additional holes from the p -doped region on the left into that same region, as in Fig. 8.7. Especially in a direct gap material, we can expect at least some of these electrons and holes to recombine, emitting photons.

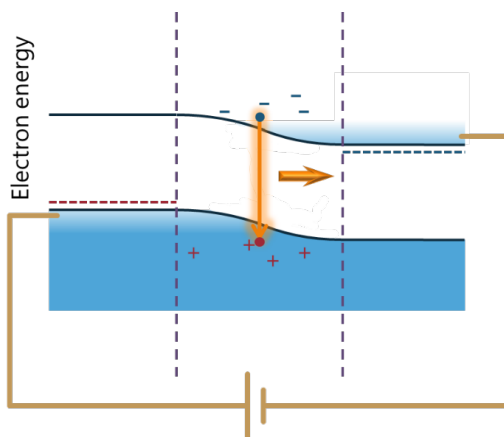


Fig. 8.7. Forward biasing a diode injects both electrons and holes into the junction region, where some of them may recombine to emit photons.

⁴ Such transitions typically have to involve some other interaction as well, with indirect transitions usually involving a crystal vibration or a “phonon” that effectively contributes the necessary crystal momentum to allow the transition to happen.

Under such forward bias, electrons and holes near the junction can effectively have their own Fermi levels, different from each other. These are known as “quasi Fermi levels”. The electrons can be in approximate thermal and diffusive equilibrium with each other, so it is meaningful to define a (quasi) Fermi level for that population of electrons. Similarly, the holes can be in approximate thermal and diffusive equilibrium with each other, with their own (quasi) Fermi level. Hence we can have useful models for their distributions. This situation is illustrated in Fig. 8.8.

Such a situation is like one we might encounter in chemistry, with two different “species” – here the electrons and the holes – each having their own chemical potentials. These two species can then “react” to generate a new species, which in the present case is the photons, with the “reaction” being the process we call “radiative recombination”.

For light-emitting diodes to begin to work, it is enough just to get some electrons in the conduction band and some holes in the valence band so that the radiative recombination can happen. Indeed, forward biasing a diode made from direct gap materials will quite generally result in light emission by this process. When we design such device well, such light-emitting diodes can be very efficient sources of light. Generally, the emission of light is at photon energies close to, but above, the bandgap energy of the semiconductor because the holes gather near to the top of the valence band and the electrons gather near to the bottom of the conduction band.

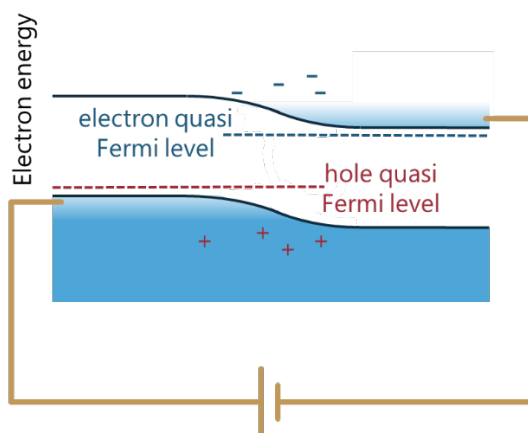


Fig. 8.8. Illustration of the quasi Fermi levels for the electrons and for the holes inside the junction region in a forward-biased diode.

Absorption and emission processes 25a

In our discussion in Chapter 7 of absorption and emission of light, we noted that there are three processes: 1) absorption of a photon to take an atom or other quantum system from a lower state to an upper state; 2) spontaneous emission in which an atom or other quantum system falls from an upper state to a lower state, emitting a photon; and 3) stimulated emission. In stimulated emission, just as for spontaneous emission, the atom or atomic system starts in an upper state, but now there are already some photons in a given mode. In spontaneous emission, the emitted photon is into the same mode as the existing ones. Since that stimulated emission process is proportional to the number of

photons already in that mode, this can give amplification of the number of photons in that mode.

This stimulated emission process exists all the time in our environment, but because in natural light there are generally very few photons in each mode (in fact, much less than one), this process is usually relatively negligible. However, it forms the basis for the laser, in which case the stimulated emission dominates, resulting in laser beams that can be tightly concentrated around one direction (or, equivalently, with most of the photons in the same mode).

Lasers 25b

Lasers are light sources in which the dominant emission of light is by stimulated emission. The word “laser” is an acronym for Light Amplification by Stimulated Emission of Radiation, and the acronym is a variant on the earlier “maser” that generated microwave radiation.

There are many different kinds of lasers. The first used specific transitions in ruby in 1960. Various approaches for lasers use transitions (a) associated with specific impurities in glasses or crystals, (b) between atomic or molecular levels in gasses, (c) in dyes in liquid form, and (d) in semiconductor diodes. Following the first demonstration in 1970 of semiconductor diode lasers that could operate continuously, the use of semiconductor diode lasers grew particularly strongly. Now, especially because of their extensive use in optical communications, diode lasers dominate over all other approaches in terms of the number in use.

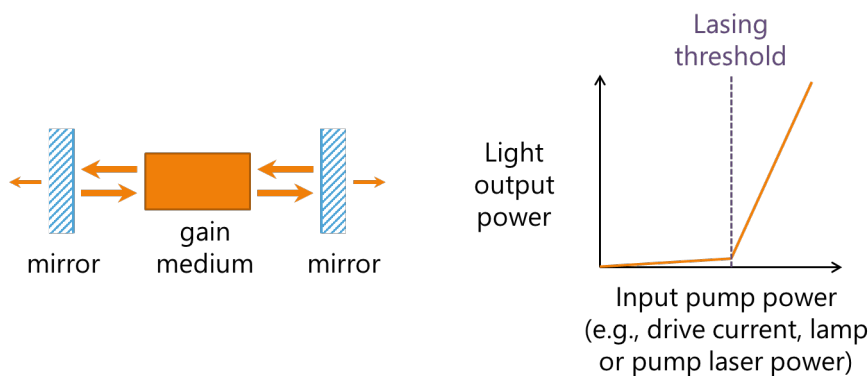


Fig. 8.9. Left: Sketch of a laser, consisting of a gain medium and mirrors (possibly partially transmitting) to reflect many of the photons back into the gain medium. Right: Sketch of the threshold behavior of the laser light output as a function of the “pumping” of the laser gain medium

A laser generally consists of a “gain medium” and some sort of resonator with mirrors, as sketched on the left in Fig. 8.9. The concept here is first that, because of stimulated emission, more photons come out of the gain medium in a given electromagnetic mode than went in. Then mirrors, forming a “cavity”, reflect the photons in that mode back

and forth through the gain medium to generate more such photons by more stimulated emission, and so on.

We may allow some photons to be transmitted through one or both mirrors so we get some optical output power. We design the device, however, so that the gain of photons from the stimulated emission in the gain medium exceeds the loss of photons both by transmission through the mirror(s) and by other absorption or scattering processes. Then the “lasing” process can grow regeneratively to give a large number of photons circulating (and emitted) in one mode, which is the laser action.

Generally, lasers have a “threshold” level of drive or “pumping” below which they will not lase. This is typically because there will be some background fraction of photons that will always be lost due to mirror transmission and other losses, and the gain per pass through the cavity has to rise to a level above that loss before the regenerative “lasing” process can take over. Below that threshold, there may still be light emission, predominantly through conventional spontaneous emission, known as “luminescence”; that spontaneous emission may happen into all the modes of the electromagnetic field at appropriate frequencies corresponding to emission transitions in the gain medium. Above the threshold for lasing, the emission into the lasing mode will tend to take over as the dominant light emission. This “threshold” emission behavior is sketched on the right of Fig. 8.9.

The laser will tend to lase in the “strongest” mode – that is, the mode with the largest positive difference between gain and loss per “round trip” in the cavity. This lasing mode can correspond to a very specific wavelength, which may correspond with the wavelength with the largest “gain” in the gain medium. The shape of the beam can also be quite a narrow beam in a very specific direction, with the precise shape of the laser beam being set by the form of the specific “strongest” mode of the resonator formed by the mirrors. (Typically, the “strongest” mode is the one with the lowest loss, and, more generally, the one with the greatest positive difference between gain and loss.)

Population inversion

A necessary condition for a laser to work is that the probability of stimulated emission in a given transition from a high energy level 2 of some quantum system to a low energy level 1 of that quantum system has to exceed the probability of absorption, which is the transition from the low energy state 1 to the high energy state 2. Otherwise, photons coming in are more likely to be absorbed than to stimulate emission, and there will be no “gain” – that is we will have fewer photons coming out of the gain medium than went in.

Now, we know from Einstein’s A and B coefficient argument that, for N_2 such quantum systems in the upper state 2, the stimulated emission rate is $B_{12}N_2\rho(\nu)$ and, for N_1 such quantum systems in the lower state, the absorption rate is $B_{12}N_1\rho(\nu)$; note that we have used Einstein’s conclusion that the same microscopic coefficient (Einstein’s B_{12} coefficient) governs both processes. So the condition for there to be a greater probability of stimulated emission than absorption is that

$$N_2 > N_1 \quad (8.2)$$

This requirement is called an “inverted population”. One reason why we do not see much stimulated emission in everyday life is that no system in thermal equilibrium will have the population of the higher energy state of the system more likely than the population of the lower energy state of the system; we can see this from Boltzmann’s factor, for example. So no thermal light source will ever give us population inversion, and so, for any given photon entering such a system at thermal equilibrium, it is always more likely to be absorbed than it is to stimulate the emission of another photon⁵.

Suppose, however, by some scheme of pumping atoms or quantum systems into their upper states (so increasing N_2), and, if necessary clearing out the “lower” states (so reducing N_1), we can “invert” the population. So, instead of the set of “atoms” or the material having net loss for photons, it can have net gain. We could express this gain as a factor G , the ratio of photons out to photons in, or as the gain γ per unit length, relating the two by

$$G = \exp(\gamma L) \quad (8.3)$$

for a gain region of length L . For example, we commonly express γ in “per centimeter”, cm^{-1} or “inverse centimeters”.

So, for example, as illustrated in the left of Fig. 8.10, we could “pump” a system somehow from a level 1 to a level 3, then have it “relax” to level 2. If the “pumping” and the relaxation are fast enough, then we may be able to make it more likely that the system is in level 2 than it is in level 1, and so we could invert the population of such “atoms”.

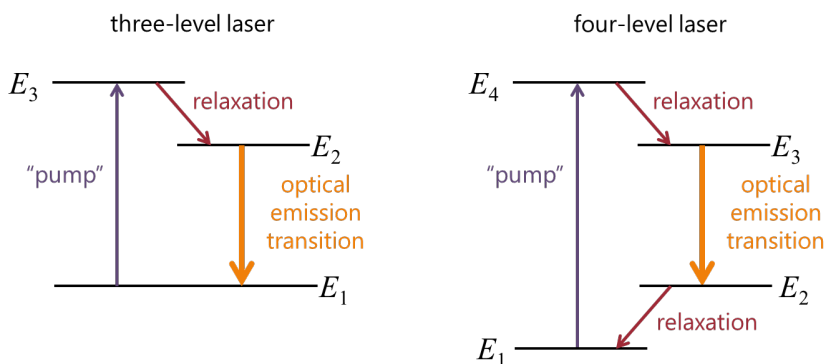


Fig. 8.10. Three-level (left) and four-level (right) laser level schemes.

There are various ways of accomplishing the “pumping”, including electrical discharge to excite the atoms or molecules, or optical pumping, such as with a flash-lamp or even another laser. Typically, the “relaxation” will be some non-radiative process, such as

⁵ Note that we are not saying that the population has to be inverted before we see stimulated emission. Stimulated emission can and will certainly occur if we have some systems in their upper state and have incoming photons of the right photon energy to match the energy difference between the states. However, if the population is not inverted, we are more likely to see absorption from the lower state to the higher state rather than stimulated emission from the higher state to the lower state.

collisions or some other scattering process. Radiative processes can be problematic for the relaxation since the emitted photons could then excite other “atoms” from level 2 back to level 3, which is undesirable when we are trying to maximize the population of level 2.

Such a three-level system can work for lasers. It has the challenge, though, that we have to make sure we excite a majority of all the “atoms” in this way; otherwise, we cannot invert the population – any “atoms” left in their ground state, level 1, are able to absorb light. The four-level system as shown in Fig. 8.10 on the right helps avoid this problem. Now we are only trying to invert the population between levels 2 and 3. If some of the “atoms” are left in level 1 because they are not excited at all, this does not matter for the population inversion of levels 2 and 3. Hence, we are not required to excite a majority of the “atoms”, which opens up more possibilities for viable lasers. In the four-level scheme, it is important that the relaxation from level 2 to level 1 is relatively rapid, so the lower “lasing level”, level 2, is kept as nearly empty as possible; this relaxation should ideally be non-radiative also.

Semiconductor lasers

Semiconductor lasers are effectively four-level lasers. Using a semiconductor diode in forward bias, just as discussed for light-emitting diodes in general, we pump electrons and holes into the junction region. If we pump hard enough, we may manage to get an inverted population. A simple version of such an inverted population would be for the (quasi) Fermi levels of both electrons and holes to move into the corresponding bands, as sketched in Fig. 8.11, which would give “pools” of electrons in the conduction band and holes in the valence band. Such a situation would certainly be inverted, because at a $\mathbf{k}$ state where we find an electron in the conduction band, we find a hole in the valence band.

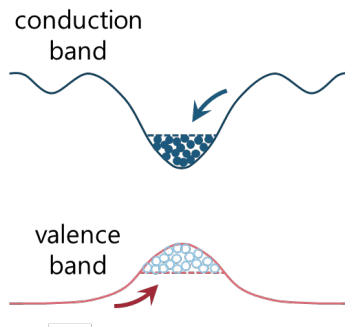


Fig. 8.11. Sketch of an inverted population in semiconductor bands, with a “pool” of electrons in the conduction band and a “pool” of holes in the valence band, giving an inverted population.

The conduction electron population at lower energies inside the pool is continually repopulated by relaxation from higher-energy states, with the whole population being replenished by the forward current. Similarly, electrons that fall into states in the top of the valence band after emitting a photon then tend to be “cleared out” of such states by holes relaxing “upwards” into these states.

The situation need not be quite as extreme as shown in Fig. 8.11 to achieve effective inverted populations in semiconductors⁶. Indeed, because of their larger effective mass the holes may well not have the Fermi level in the valence band anyway, and in practice the “pools” do not usually have such well-defined “surfaces”, having instead widths of the order of a few $k_B T$ because of the Fermi-Dirac distribution.

Semiconductor lasers can have particularly large “gain”, which makes them relatively easy to make. They can also be very compact, with size scales as small as microns to hundreds of microns in many types in current use. Among lasers that are directly driven electrically, they are also particularly efficient.

Quiz questions

- Q8.3.1 If I heat up some material sufficiently, I will be able to get it to “lase” without any other source of power. True or False?
- Q8.3.2 Because absorption of a photon is more likely than emission of a photon, then when I shine light on some warm body, the warm body cannot emit light. True or False?
- Q8.3.3 Which kind of semiconductor material would you choose if you wanted to make a good light emitter - indirect gap or direct gap?
- Q8.3.4 If I forward bias a diode, I will only get light emission if the diode heats up to a high temperature. True or False?
- Q8.3.5 In a forward biased light-emitting diode, the electrons and holes move into the same region as a result primarily of diffusion. True or False?
- Q8.3.6 Absorption is the process of taking a system in a high-energy state 1 and changing it to being in a low-energy state 2 by absorbing a photon. True or False?
- Q8.3.7 Just as there is a process of spontaneous (photon) emission, in which a system in a high-energy state can fall to a low-energy state by emitting a photon, even if there are no other photons present, there is a process of spontaneous (photon)

⁶ In a semiconductor, for an electron state in the conduction band and another one at essentially the same $\mathbf{k}$ in the valence band, the probability of a photon causing a radiative (absorption or emission) transition from an electron state in one band to an electron state in the other is proportional to the probability f_e that the starting state is occupied by an electron (or the probability $1 - f_h$ that it is “empty” of a hole) and the probability $1 - f_e$ that the ending state is empty of an electron (or the probability f_h that it is “full” with a hole). Here we use the f_e Fermi-Dirac distribution for electrons in the conduction band, with the electron quasi Fermi level, and the f_h Fermi-Dirac distribution for holes in the valence band, with the hole quasi Fermi level. Hence, the probability of absorption on this transition is $\propto (1 - f_e)(1 - f_h)$ and the probability of (stimulated) emission on this transition is $\propto f_e f_h$. The difference between these two probabilities is therefore $\propto f_e f_h - (1 - f_e)(1 - f_h) = (f_e + f_h) - 1$. So stimulated emission is more likely than absorption for $f_e + f_h > 1$, which becomes the effective condition for an inverted population for a given $\mathbf{k}$ state in a semiconductor.

absorption in which a system in a low-energy state can transition to being in a high energy state, even if there are no photons present. True or False?

Q8.3.8 Light from a laser is always stimulated emission. True or False?

Q8.3.9 For a laser to work, it must be heated to a high temperature to get sufficient populations in the upper state of the lasing transition. True or False?

Q8.3.10 To make a laser, it is sufficient to get the gain material to show some gain (i.e., more photons out in some mode than the number of photons we put in). True or False?

Q8.3.11 If a laser is not working, I should turn it upside-down to invert the population. True or False?

Problem

8.3.1 I have two different light-emitting diodes (LEDs). One is designed to emit blue light, and the other is designed to emit red light. For the purposes of this question, you can presume that the LEDs all emit light with photon energy very close to the bandgap energy of the semiconductor material from which they are made (and, therefore, we are using different semiconductors for the two different colors of LEDs).

In this experiment, we will be shining light from one LED into another, and looking at consequences for the LED into which I am shining the light. Note also that LEDs, since they are diodes, can also function as photodiode detectors if we shine appropriate light into them.

- (i) Suppose first of all, that I shine light into one of the LEDs, and I see current in a current meter connected to the LED. What is the sign of the electric current that I see? Specifically, if I connect the p end of the LED to the “+” terminal of the current meter and the n end of the LED to the “-” negative terminal of the current meter, will it show a positive or negative current? (A positive electrical current flowing into the “+” terminal of the meter will be a positive current in this sense).
- (ii) I connect a power supply to the blue LED, forward biasing it, and as a result it emits blue light. I shine this light into the red LED. The red LED is connected just to a current meter. Will I expect to see significant current in the current meter?
- (iii) As in (ii) above, I bias the blue LED so it is emitting blue light, and shine that into the red LED. This time the red LED is connected just to a volt meter. Will I expect to see significant voltage on the volt meter?
- (iv) I flip round the experiment from part (ii), connecting a power supply to the red LED to forward bias it, as a result of which it emits red light. I shine this red light into the blue LED. The blue LED is connected just to a current meter. Will I expect to see significant current in the current meter?

- (v) As in (ii) above, I shine light from the blue LED into the red LED, but I do not connect the red LED terminals to anything at all. Will I be able to see red light emitted from the red LED if I look carefully?

8.4 Conclusions

The use of semiconductors, and especially semiconductor diode structures, opens up a wide range of useful optoelectronic devices, with many useful applications. Such structures allow us to generate solar power. The use of light-emitting diodes for illumination may well also save substantial amounts of power compared to previous methods. Semiconductor lasers and other devices like semiconductor optical modulators transmit nearly all of our data over optical fibers, and semiconductor photodetectors receive it at the other end. Semiconductor detectors now record nearly all of our pictures in digital cameras. Despite this wide range of devices and applications, the underlying principles of all these devices are largely covered by the physics we have introduced here, and that physics lays a solid foundation for the engineering of optoelectronic devices and systems.

Chapter 9

Epilogue

Now that we have reached the end of this story, we can look back at how we progressed through a linked set of ideas. We started with modes and eigenstates in the classical world of oscillations and waves, ideas that are quite useful there in themselves. These ideas prepared us for the quantum mechanical view: eigenstates become the way we look even at states of matter, and form a surprising and different basis for light and electromagnetism. Armed with these ideas of states of matter and light, we introduced the statistical view of the occupation of these states. That statistical view gives us the idea of entropy. Entropy then forms the core of the whole fields of statistical mechanics and thermodynamics. Those fields then explain much of the macroscopic behavior of the world around us while also linking back to the microscopic quantum mechanics.

If you came to this material just with a background such as basic classical mechanics and electromagnetism, I think you now understand that background on its own does not come close to explaining the world around us; it also does not give us many of the tools that are essential for constructing the technology that enables much of our modern lives. But I also hope you now understand that there is a relatively small number of key ideas that do allow us to explain how the everyday world around us actually works.

We have seen that often this explanation is counter to our intuition. The quantum mechanical view necessarily has many concepts in it that do not agree with our everyday “common sense”, but that nonetheless do give a coherent and practically useful basis for explaining and engineering the world. We do also have to be honest that there are a few aspects of the basics of quantum mechanics that still puzzle us, especially the idea of quantum mechanical measurement. We learn to live with those, however, as we use the practical techniques of quantum mechanics to design real materials, devices and technology.

The ideas of statistical mechanics explain much about complex systems and why, because of overwhelming statistical probabilities, much of what is random can have apparently coherent and predictable behavior. If we understand that the larger system is always, in a statistical sense, going to proceed whenever possible to the macrostate with

the largest multiplicity or entropy, then we have the key principle that guides many of the processes we see around us.

In the space available here, we could only just introduce many of the core concepts, often with simple “toy” models, though sometimes with quite practical extensions through to the principles of modern devices and to behaviors we see in the macroscopic world every day. Of course, there is always more to learn about the topics we introduced in the preceding chapters. But you can be sure that the material and ideas we have covered here are key physical ideas that do largely explain the everyday modern world around us; they give us both the concepts and the vocabulary to proceed to the next level, both in science and engineering.

If this is where you have to leave this story, at least for the present, then you do so much better informed about a physical world that both enables and bounds much of what we can do. If you are heading into a deeper study or use of physical science or technology, then you can have the confidence that you now understand a core set of concepts that is quite finite, reliable and universally useful, and that will give a solid foundation for all that you are about to learn there. Whichever path you are taking, I hope that you have found this story here at least informative, possibly interesting, and hopefully even fun!

Appendix A

Physical constants and units

A.1 Fundamental constants

Constant Name	Symbol	Numerical Value	Δ	Units
Bohr magneton	μ_B	$9.274\,009\,68 \times 10^{-24}$	20	J T ⁻¹
Boltzmann constant	k_B	$1.380\,6488 \times 10^{-23}$	13	J K ⁻¹
Electric constant	ϵ_o	$8.854\,187\,817\dots \times 10^{-12}$	-	F m ⁻¹
Electron g factor	g_e	2.002 319 304 361 53	53	-
Electron mass	m_e	$9.109\,382\,91 \times 10^{-31}$	40	kg
Elementary charge	e	$1.602\,176\,565 \times 10^{-19}$	35	C
Fine structure constant	α	$7.297\,352\,5698 \times 10^{-3}$	24	-
Magnetic constant	μ_o	$4\pi \times 10^{-7}$	-	H m ⁻¹
Planck's constant	h	$6.626\,069\,57 \times 10^{-34}$	29	J s
Planck's constant/ 2π	$\hbar$	$1.054\,571\,726 \times 10^{-34}$	47	J s
Proton mass	m_p	$1.672\,621\,777 \times 10^{-27}$	74	kg
Proton-electron mass ratio	m_p / m_e	1 836.152 672 45	75	-
Rydberg constant	R_∞	10 973 731.568 539	55	m ⁻¹
	$R_\infty hc / e$	13.605 692 53	30	eV
Speed of light in vacuum	c	299 792 458	-	m s ⁻¹

The “ Δ ” quoted is the absolute value of the uncertainty in the last two digits of the quoted numerical value corresponding to one standard deviation from the numerical value given. So, for example, the possible values of Planck's constant within one standard deviation of the best estimate lie between 6.626 069 28 and 6.626 069 86 J s.

The speed of light in vacuum has been chosen to have the exact value shown because the meter is now defined as the length of the path traveled by light in vacuum during the time interval of 1/299 792 458 of a second. The magnetic constant (also known as the permeability of free space) is chosen to have the value shown because it is an arbitrary constant that arises from the choice of the system of units and the electric constant (also known as the permittivity of free space) then follows from it and the (chosen) velocity of light because, by definition, $c = 1/\sqrt{\epsilon_o \mu_o}$, so all three of these

quantities have no uncertainty by definition. The Bohr magneton is $\mu_B = e\hbar/2m_e$. The fine structure constant is $\alpha = e^2/4\pi\epsilon_0 c\hbar$.

These values are the CODATA Internationally recommended values as of 2010. Reference <http://physics.nist.gov/cuu/Constants/index.html>.

A.2 SI units

We list here most of the major SI base and derived units. For a full list, see <http://physics.nist.gov/cuu/Units/units.html>.

SI base units

Base quantity	Name	Symbol
length	meter	m
mass	kilogram	kg
time	second	s
electric current	ampere	A
thermodynamic temperature	kelvin	K

SI derived units

Derived quantity	Name	Symbol	In terms of other SI units	In terms of SI base units
frequency	hertz	Hz	-	s ⁻¹
force	newton	N	-	m·kg·s ⁻²
pressure, stress	pascal	Pa	N/m ²	m ⁻¹ ·kg·s ⁻²
energy, work, quantity of heat	joule	J	N·m	m ² ·kg·s ⁻²
power, radiant flux	watt	W	J/s	m ² ·kg·s ⁻³
electric charge, quantity of electricity	coulomb	C	-	s·A
electric potential difference, electromotive force	volt	V	W/A	m ² ·kg·s ⁻³ ·A ⁻¹
capacitance	farad	F	C/V	m ⁻² ·kg ⁻¹ ·s ⁴ ·A ²
electric resistance	ohm		V/A	m ² ·kg·s ⁻³ ·A ⁻²
electric conductance	siemens	S	A/V	m ⁻² ·kg ⁻¹ ·s ³ ·A ²
magnetic flux	weber	Wb	V·s	m ² ·kg·s ⁻² ·A ⁻¹
magnetic flux density	tesla	T	Wb/m ²	kg·s ⁻² ·A ⁻¹
inductance	henry	H	Wb/A	m ² ·kg·s ⁻² ·A ⁻²

SI Prefixes

Factor	Name	Symbol
10^{24}	yotta	Y
10^{21}	zetta	Z
10^{18}	exa	E
10^{15}	peta	P
10^{12}	tera	T
10^9	giga	G
10^6	mega	M
103	kilo	k
10^2	hecto	h
10^1	deka	da

Factor	Name	Symbol
10^{-1}	deci	d
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	μ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a
10^{-21}	zepto	z
10^{-24}	yocto	y

Appendix B

Greek alphabet

Upper Case	Lower Case	Name
A	α	alpha
B	β	beta
Γ	γ	gamma
Δ	δ	delta
E	ε	epsilon
Z	ζ	zeta
H	η	eta
Θ	θ	theta
I	ι	iota
K	κ	kappa
Λ	λ	lambda
M	μ	mu
N	ν	nu
Ξ	ξ	xi
O	ο	omicron
Π	π	pi
P	ρ	rho
Σ	σ	sigma
T	τ	tau
Υ	υ	upsilon
Φ	φ	phi
X	χ	chi
Ψ	ψ	psi
Ω	ω	omega

Appendix C

Thermal carrier distributions

C.1 Introduction

In this Appendix, we extend the qualitative discussion in Chapter 6 of Fermi levels, doping, and electron and hole populations in semiconductors, showing how simple quantitative relations can be established. The main additional concept we need to introduce is the idea of the density of states, which lets us model how many electron (or hole) states there are within given energy ranges. Together with the Fermi-Dirac distribution, we can then connect carrier (electron and/or hole) populations quantitatively to Fermi levels.

C.2 Density of states

To understand quantitatively how electrons are distributed among the available states, we need to know how the states themselves are distributed in energy. In particular, we would like to know how many electron states there are within some small energy range ΔE in the vicinity of some electron energy E . This number of states possible per unit energy range is known as the density of states (in energy).

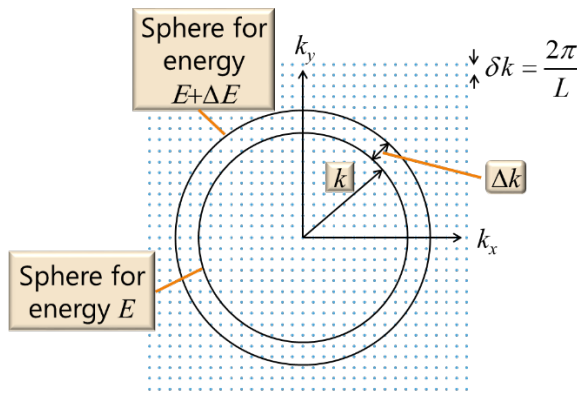


Fig. C.1. Sketch of equally spaced points in k_x and k_y and spheres of constant energy for the case of an isotropic parabolic band.

Now, we know from the Bloch theorem that the allowed values of k are separated by an amount $2\pi / Na$ for a one-dimensional “crystal” with N unit cells of size a . Equivalently, if we said the crystal was of length $L = Na$, then the allowed values of k are separated by $2\pi / L$. Now we generalize to three dimensions. For simplicity, we will presume the crystal is a cube of side L . So now we presume for a wavevector $\mathbf{k}$ that each of its components – k_x , k_y , and k_z in the x , y , and z directions respectively – can take on values similarly spaced by Δk where

$$\delta k = \frac{2\pi}{L} \quad (\text{C.1})$$

We could sketch out all of the allowed values of $\mathbf{k}$ as a set of points in a three dimensional diagram plotted as a function of the components k_x , k_y , and k_z . This would just be a regular set of points equally spaced in all three directions. In Fig. C.1, we have sketched this set of points, though we are showing only a two-dimensional cross-section in the k_x and k_y plane at $k_z = 0$.

Though we know how the various possible states are spaced in $\mathbf{k}$, we need to know how they are spaced in energy E . To do so, we need some model – a band structure – that relates E and $\mathbf{k}$. We saw that, near to a band minimum or maximum, we expect parabolic relations between E and components of $\mathbf{k}$. The simplest situation to consider is where this parabolic relation is the same in all three directions, which is what we would call an isotropic¹ parabolic “band” (or region of the band structure). Such a model is a good approximation for many bands near minima or maxima². Taking the energy at the minimum to be zero for simplicity, such a relation corresponds to

$$E = \frac{\hbar^2 k^2}{2m_{\text{eff}}} \quad (\text{C.2})$$

where the effective mass m_{eff} is parametrizing the curvature of the parabola. So, for some specific energy E , we could draw a corresponding sphere in our diagram Fig. C.1, of radius

$$k = \sqrt{\frac{2m_{\text{eff}} E}{\hbar^2}} \quad (\text{C.3})$$

For some slightly larger energy $E + \Delta E$ we could draw a sphere of a slightly larger radius $k + \Delta k$. From our relation Eq. (C.2), we know how to relate Δk and ΔE .

¹ “isotropic” means the same in all directions (from the Greek “iso” meaning “the same” and “tropos” meaning a turning, way or manner).

² This isotropic parabolic band approach is a good first approximation near zone center for the valence bands in many semiconductors, including Si, Ge, and the III-V materials that crystallize in zinc blende form, and for the lowest conduction band in zinc blende direct gap III-V materials like GaAs or InP. For the lowest conduction bands in Si and Ge, in part because their lowest points do not lie at zone center, the band minima, though approximately parabolic in each direction, have one effective mass in one direction and a different one in the other two. The approach we are taking here can be extended to such cases, though we will not do that here.

$$k + \Delta k = \sqrt{\frac{2m_{\text{eff}}(E + \Delta E)}{\hbar^2}} = \sqrt{\frac{2m_{\text{eff}}E}{\hbar^2}} \sqrt{1 + \frac{\Delta E}{E}} \approx \sqrt{\frac{2m_{\text{eff}}E}{\hbar^2}} \left(1 + \frac{\Delta E}{2E}\right) \quad (\text{C.4})$$

where we used the expansion for small x of $\sqrt{1+x} \approx 1 + x/2$. So, subtracting Eq. (C.3) from Eq. (C.4) gives

$$\Delta k \approx \sqrt{\frac{2m_{\text{eff}}E}{\hbar^2}} \frac{\Delta E}{2E} \quad (\text{C.5})$$

Now, the number of allowed $\mathbf{k}$ states that lie between energy E and energy $E + \Delta E$ is just the number of “dots” in Fig. C.1, extended into three dimensions, that lie between the sphere for E and the sphere for $E + \Delta E$. On this figure, the “volume” of that space is just the “volume” of a spherical shell of (inner) radius k and thickness Δk . The surface “area” of a sphere of radius k is $4\pi k^2$. So the “volume” of a thin spherical shell with thickness Δk is (for small Δk) $4\pi k^2 \Delta k$. We can think of each “dot” in Fig. C.1 as being surrounded by a cubic “volume” of magnitude $(2\pi/L)^3$, so the number of such “dots” that lie in this spherical shell is, using Eqs. (C.2), (C.3) and (C.5),

$$\frac{4\pi k^2 \Delta k}{(2\pi/L)^3} = \frac{L^3}{2\pi^2} k^2 \Delta k = \frac{L^3}{2\pi^2} \frac{2m_{\text{eff}}E}{\hbar^2} \sqrt{\frac{2m_{\text{eff}}E}{\hbar^2}} \frac{\Delta E}{2E} = \frac{L^3}{4\pi^2} \left(\frac{2m_{\text{eff}}}{\hbar^2}\right)^{3/2} E^{1/2} \Delta E \quad (\text{C.6})$$

This is the number of $\mathbf{k}$ states lying between energy E and energy $E + \Delta E$. To get the total number of electron states in this energy range (presuming we have no magnetic field to give different energies to different spins), we multiply by 2 for the 2 different spins. We note that L^3 is just the volume of the crystal. So we can also divide by L^3 and write that the total number of electron states between E and $E + \Delta E$, per unit volume of crystal, is

$$g(E) \Delta E = \frac{1}{2\pi^2} \left(\frac{2m_{\text{eff}}}{\hbar^2}\right)^{3/2} E^{1/2} \Delta E \quad (\text{C.7})$$

This quantity

$$g(E) = \frac{1}{2\pi^2} \left(\frac{2m_{\text{eff}}}{\hbar^2}\right)^{3/2} E^{1/2} \quad (\text{C.8})$$

is known as the density of states in energy (per unit volume of crystal) for an isotropic parabolic band. With this quantity and the Fermi-Dirac distribution, we can start to calculate useful distributions of electrons in such materials.

C.3 Carrier concentrations and Fermi levels

Suppose we know the Fermi level E_F in some crystalline material. We presume also that we know $g_C(E - E_C)$, the density of states in energy (per unit volume of crystal) for electrons in the conduction band of interest, where the lowest energy in the band is E_C ; for example, this density of states might have the simple isotropic parabolic band form of Eq. (C.8) for some electron effective mass m_e .

Then we can write down an expression for the total number of electrons in this conduction band (per unit volume of crystal). This will be the integral of the density of states multiplied by the Fermi-Dirac distribution function that tells us the probability that any given state is occupied in thermal equilibrium. So we have, for the total number N_e of electrons (per unit volume of crystal)

$$N_e = \int_{E_C}^{\infty} g_C(E - E_C) f_e(E) dE \quad (C.9)$$

where f_e is the Fermi-Dirac distribution function for these electrons, that is

$$f_e(E) = \frac{1}{\exp\left(\frac{E - E_F}{k_B T}\right) + 1} \quad (C.10)$$

Now we can calculate results for the total number of electrons in bands for some specific cases.

Generally, it is relatively straightforward mathematically to calculate the number of electrons N_e if we know the Fermi level and the form of the density states – for example, from our knowledge of the band structure; we simply have to integrate Eq. (C.9). It is, however, often harder to deduce the Fermi level from a knowledge of the number of electrons because there is often no analytic formula for the integral in Eq. (C.9) for even simple forms of the density of states. Hence, we often cannot simply algebraically invert such a formula to deduce the Fermi level, though we can always do so numerically or with some numerical approximation.

There are, however, at least two useful limiting cases, known as “degenerate”³ and “non-degenerate”, in which we can establish simple relations, and these also correspond to common physical situations, so we will examine them below.

C.4 Metals and degenerate statistics

For a metal or for some other situation where the Fermi level is many $k_B T$ above the bottom of a band, as a first approximation in evaluating the integral Eq. (C.9) for the total number of electrons, we can approximate the Fermi-Dirac distribution by its version at zero temperature; in that case the Fermi-Dirac distribution is 1 up to the Fermi level and 0 above it, so the integral Eq. (C.9) becomes

³ This use of “degenerate” is, unfortunately, quite different from the use of the term “degeneracy” to refer to the number of quantum mechanical states that have the same energy in some system (such as the 2s and 2p states in hydrogen, which all have the same energy in the simplest model of hydrogen), or more generally to describe the situation of multiple different eigenfunctions for the same eigenvalue. There is no specific relation between these two different uses of the same word.

$$\begin{aligned}
 N_e &= \int_{E_C}^{E_F} g_C(E - E_C) dE = \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \int_{E_C}^{E_F} (E - E_C)^{1/2} dE \\
 &= \frac{1}{3\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} (E_F - E_C)^{3/2}
 \end{aligned}
 \tag{C.11}$$

We can easily invert the expression Eq. (C.11) to obtain the Fermi energy as a function of the total number (density) N_e of electrons in the band, giving

$$E_F - E_C = \frac{\hbar^2}{2m_e} (3\pi^2 N_e)^{2/3} \tag{C.12}$$

Incidentally, a useful “order-of-magnitude” number to have in mind when considering the statistics of electrons in materials (and for other purposes as well) is the size of $k_B T$ at room temperature, in electron-volts. For a typical room temperature of, say, 20° C (so 293.15 K), using $k_B T$ in electron-volts as $k_B T$ in joules divided by the magnitude e of the electronic charge, we have⁴

$$k_B T \text{ at } 20^\circ \text{ C} \cong 25 \text{ meV} \tag{C.13}$$

Metals typically have Fermi levels that are of the order of electron-volts relative to the bottom of their conduction band, so the “zero-temperature” approximation is very good since $k_B T$ is a very small fraction of the Fermi level energy relative to the bottom of the band.

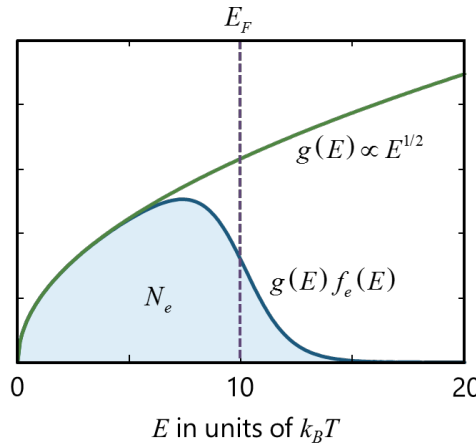


Fig. C.2. Plot of $g(E)$, the density of states in energy for a parabolic band, and $g(E)f_e(E)$, the product of this density of states and the Fermi-Dirac function f_e , for the case where the Fermi energy $E_F = 10k_B T$. The electron energy E is plotted in units of $k_B T$ from the bottom of the band. The shaded area represents the total number of electrons, N_e , which is the integral of $g(E)f_e(E)$ over all energies.

⁴ A more exact value for 20° C would be 25.2671 meV.

A more intermediate situation is sketched in Fig. C.2, in which the Fermi level is $10k_B T$ above the bottom of the band. This figure also illustrates the form of the density of states for a parabolic band, and the product of the density of states and the Fermi-Dirac distribution – that is, the function $g(E)f_e(E)$ – which is the form of the integrand in Eq. (C.9). The shaded area in Fig. C.2 represents the result of the integral – that is, the total number (density) N_e of electrons in the band – in Eq. (C.9).

A situation where the Fermi level is many $k_B T$ above the bottom of that band – as it is in a metal or in the intermediate situation of Fig. C.2 – is referred to as “degenerate” or corresponding to “degenerate” statistics, and the formula Eq. (C.12) can often be used to deduce the Fermi level.

C.5 Electrons and holes

From our discussions of the Boltzmann factor, or the Gibbs factor, or the full form of the Fermi-Dirac distribution, we know that, at finite temperature, for two quantum states separated by some energy, there is a finite probability that the system will be in the higher energy state. We remember that this finite probability comes from the need to maximize entropy overall for the system connected to a thermal reservoir.

So, when we consider a material like a semiconductor or an insulator, for any finite band gap energy and for any finite temperature, there is some possibility that there will be some electrons in quantum states in the conduction band. Because pure semiconductors and insulators have numbers of electrons that would exactly fill a band, if there are some electrons in quantum states in the conduction band, there must be some quantum states for electrons in the valence band that are empty. As discussed in Chapters 4 and 6, we call these empty states “holes”. This situation is illustrated in Fig. C.3 (see also Fig. 6.1) .

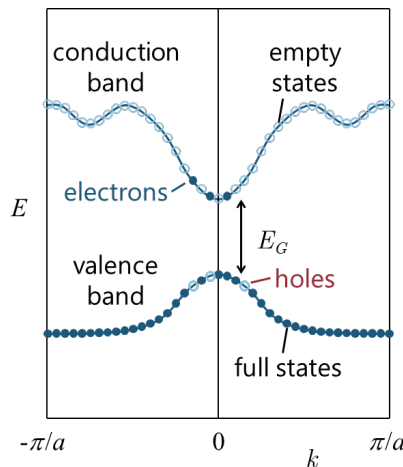


Fig. C.3. Conduction and valence bands in a semiconductor or insulator showing finite thermal excitation of electrons from the valence band to the conduction band, leaving “holes” in the valence band.

Now, we know that quite generally for the occupation of the electron states that, in equilibrium, they obey the Fermi-Dirac distribution f_e as given by Eq. (C.10). This distribution applies to all the quantum states, both in the valence band and the conduction band. Since f_e is the probability that a given quantum state of a given energy is occupied by an electron, then

$$f_h(E) = 1 - f_e(E) \quad (\text{C.14})$$

is the probability that quantum state is not occupied or is “empty”. We could also say that this is the probability that it is “occupied” or “full” with a “hole” if we regard a hole as an absence of an electron. We can if we want therefore write down a thermal distribution for such “holes”. From Eq. (C.14) we have

$$f_h(E) = 1 - \frac{1}{\exp\left(\frac{E - E_F}{k_B T}\right) + 1} = \frac{\exp\left(\frac{E - E_F}{k_B T}\right) + 1 - 1}{\exp\left(\frac{E - E_F}{k_B T}\right) + 1} = \frac{\exp\left(\frac{E - E_F}{k_B T}\right)}{\exp\left(\frac{E - E_F}{k_B T}\right) + 1} \quad (\text{C.15})$$

Multiplying the top and the bottom of Eq. (C.15) by $\exp[-(E - E_F)/k_B T]$ gives

$$f_h(E) = \frac{1}{\exp\left(-\frac{(E - E_F)}{k_B T}\right) + 1} \quad (\text{C.16})$$

This is the same form as the Fermi-Dirac distribution, but with energy E and the Fermi energy E_F regarded as minus their values in the electron distribution. What this means is that, if we “stand on our heads” and look at a figure like Fig. C.3 upside-down, we will see Fermi-Dirac distributions for holes; we simply regard all the energies for holes as being minus the corresponding energies for electrons, and we get the same form of distribution⁵.

As discussed in Chapter 4, we can view holes as being like bubbles. A liquid, which is like some “sea” of drops of liquid that are like electrons in our analogy, settles down towards the lowest energy by filling a bucket from the bottom up. Bubbles in the liquid, which are absences of drops of liquid, rise to the top of the liquid; their energies are upside down compared to those of drops of liquid – they want to fall “up”.

Incidentally, since electrons and holes can both be viewed as being able to carry electrical current, as we will discuss below, the term “carriers” is often used to cover both types when we need a generic term for that idea.

⁵ We can also view the conduction band as being mostly “filled” with holes.

C.6 Law of mass action and intrinsic carrier concentration

In pure semiconductors and insulators, the total number of electrons in the materials is such that at zero temperature, the valence bands are full and the conduction bands are empty. If we think about the Fermi-Dirac distribution, that implies that the Fermi level lies somewhere within the band gap energy range – that is, above the top (or the “edge”) of the highest-energy valence band, at an energy we will call E_V , and below the bottom (or the “edge”) of the lowest-energy conduction band, at the energy E_C (as used previously in Eq. (C.9), for example).

If we imagine that we increase the temperature, we expect that some small number of electrons will be thermally excited to the (lowest) conduction band leaving a small number of holes in the (highest) valence band. Since these numbers of electrons and holes are small, that means that the Fermi level is quite far from either band edge. In that case, we can start to analyze this problem by using the Maxwell-Boltzmann approximation to the Fermi-Dirac distributions. So, for the electrons in the conduction band, we have

$$f_e(E) = \exp\left(\frac{E_F - E_C}{k_B T}\right) \exp\left[-\frac{(E - E_C)}{k_B T}\right] \quad (\text{C.17})$$

So, using Eq. (C.9), with this approximation for f_e , and using the form of the electron density of states from Eq. (C.8) with the conduction band edge at E_C , and with electron effective mass m_{effe} , we have

$$\begin{aligned} N_e &= \frac{1}{2\pi^2} \left(\frac{2m_{effe}}{\hbar^2} \right)^{3/2} \int_{E_C}^{\infty} (E - E_C)^{1/2} \exp\left[-\frac{(E - E_C)}{k_B T}\right] dE \\ &= \frac{1}{2\pi^2} \left(\frac{2m_{effe}k_B T}{\hbar^2} \right)^{3/2} \exp\left(\frac{E_F - E_C}{k_B T}\right) \int_0^{\infty} x^{1/2} \exp(-x) dx \end{aligned} \quad (\text{C.18})$$

Since

$$\int_0^{\infty} x^{1/2} \exp(-x) dx = \frac{\sqrt{\pi}}{2} \quad (\text{C.19})$$

then Eq. (C.18) becomes, for the total number of electrons per unit volume in the conduction band,

$$N_e = N_C \exp\left(\frac{E_F - E_C}{k_B T}\right) \quad (\text{C.20})$$

where the quantity N_C , which is sometimes called the “effective density of states at the conduction band edge” is given by⁶

$$N_C = 2 \left(\frac{m_{effe} k_B T}{2\pi \hbar^2} \right)^{3/2} \approx 2.510 \times 10^{19} \left(\frac{m_{effe}}{m_e} \frac{T}{300} \right)^{3/2} \text{ cm}^{-3} \quad (\text{C.21})$$

where m_e is the mass of a free electron.

We can then repeat the same derivation, but now for the valence band. Now we choose the energy at the “top” of the valence band to be E_V . Hence we can write

$$f_h(E) = \exp\left(-\frac{E_F - E_V}{k_B T}\right) \exp\left[+\frac{(E - E_V)}{k_B T}\right]$$

as the Maxwell-Boltzmann approximation to the “hole” Fermi-Dirac function as in Eq. (C.16), and

$$g_V(E_V - E) = \frac{1}{2\pi^2} \left(\frac{2m_{effh}}{\hbar^2} \right)^{3/2} (E_V - E)^{1/2} \quad (\text{C.22})$$

where the effective mass for the holes is m_{effh} . This hole effective mass is treated as a positive number here for a maximum at the top of the valence band. The energy E here is formally the electron energy. In performing the integral here, formally we would be integrating from $-\infty$ to E_V , though mathematically we can change that to an integral from E_V to ∞ to obtain a similar integral to that of Eq. (C.18), with the result for the total number of holes (per unit volume) in the valence band being

$$N_h = N_V \exp\left(\frac{E_V - E_F}{k_B T}\right) \quad (\text{C.23})$$

where

$$N_V = 2 \left(\frac{m_{effh} k_B T}{2\pi \hbar^2} \right)^{3/2} \approx 2.510 \times 10^{19} \left(\frac{m_{effh}}{m_e} \frac{T}{300} \right)^{3/2} \text{ cm}^{-3} \quad (\text{C.24})$$

is called the “effective density of states at the valence band edge”.

Now we can multiply these two results, Eqs. (C.20) and (C.23), to obtain

$$N_e N_h = N_C N_V \exp\left(-\frac{E_G}{k_B T}\right) \quad (\text{C.25})$$

where E_G is that band gap energy, which by definition is

$$E_G = E_C - E_V \quad (\text{C.26})$$

⁶ Note that we are quoting this number in cm^{-3} , not in m^{-3} . In practice, the use of “per cubic centimeter” for specifying electron and hole densities is almost universal in semiconductor physics and devices. In m^{-3} , the numbers for densities are 10^6 larger.

Eq. (C.25) is known as the “Law of Mass Action”⁷. It tells us that, in equilibrium, at least if the Fermi energy is far from both the conduction or valence band edges, there is a reciprocal relationship between the numbers of electrons (in the conduction band) and the number of holes (in the valence band) – if we somehow increase the number of electrons by some factor, we will correspondingly decrease the number of holes by the same factor. In practice in the use of semiconductors, we make such changes by doping the semiconductor, and we will discuss doping and Fermi levels below.

In a pure semiconductor or insulator material, we know that, however many electrons are in the conduction band, that number must be balanced by the number of holes in the valence band; otherwise the material would not be electrically neutral. Hence, in the pure material, we must have

$$N_e = N_h = N_i \quad (\text{C.27})$$

where N_i is called the intrinsic⁸ carrier concentration. From the Law of Mass Action as in Eq. (C.25), we can therefore write an expression for N_i , giving

$$N_i = \sqrt{N_C N_V} \exp\left(-\frac{E_G}{2k_B T}\right) \quad (\text{C.28})$$

This is the electron and hole concentration we expect in a pure semiconductor or insulator. The factors N_C and N_V do increase with temperature, and through the different effective masses in different materials, they do vary somewhat from material to material because of different effective masses. However, the dominant factor in determining whether a pure material is a good insulator at a given temperature or whether it has some significant conductivity from the “intrinsic” presence of electrons and holes (as in a “semiconductor” material) is the band gap energy E_G in relation to $k_B T$.

For example, at about room temperature, with $k_B T \approx 25$ meV, we could consider two materials, the second with a band gap energy 1 eV larger than the first. This 1 eV difference will lead to a decrease of the exponential factor in Eq. (C.28) by $\sim \exp(-20) \approx 2 \times 10^{-9}$ in the second material compared to the first. Though we should consider the factors N_C and N_V and also consider the mobilities of the electrons and holes to get a precise answer, overall, we therefore expect the electrical conductivity in the materials to differ by a factor of roughly this order of magnitude. So, if the first material was a moderate conductor, we would expect the second material with the 1 eV larger band gap to be a reasonably good insulator.

This approach also enables us to calculate where the Fermi level is in such pure semiconductors and insulators. Since the numbers of electrons and holes must be equal in such pure materials (Eq. (C.27)), we must have, from Eqs. (C.20) and (C.23)

⁷ This name comes originally from chemistry, though the statement of this law in chemistry looks somewhat different from this form even though the underlying principles are the same.

⁸ The word “intrinsic” is used to refer to the properties of materials without added impurities or “dopants”. Such a pure material could also be described as an “intrinsic” material.

$$N_C \exp\left(\frac{E_F - E_C}{k_B T}\right) = N_V \exp\left(\frac{E_V - E_F}{k_B T}\right) \quad (\text{C.29})$$

or equivalently, using Eqs. (C.21) and (C.24)

$$\frac{N_C}{N_V} = \left(\frac{m_{effe}}{m_{effh}}\right)^{3/2} = \exp\left(\frac{E_V - E_F - E_F + E_C}{k_B T}\right) \quad (\text{C.30})$$

Taking the logarithm of both sides and multiplying both sides by $k_B T$ and using Eq. (C.26) gives

$$\frac{3k_B T}{2} \log \frac{m_{effe}}{m_{effh}} = E_V + E_C - 2E_F = 2E_V + E_G - 2E_F \quad (\text{C.31})$$

So

$$E_F = E_V + \frac{E_G}{2} - \frac{3k_B T}{4} \log \frac{m_{effe}}{m_{effh}} \quad (\text{C.32})$$

Effective masses can differ substantially between the valence and conduction bands – for example for GaAs, $m_{effe} \approx 0.07m_e$ and $m_{effh} \approx 0.4m_e$. Since only the logarithm of the ratio enters in to Eq. (C.32), however, we can still say that the Fermi level in such pure semiconductors or insulators lies within a few $k_B T$ of the middle of the band gap. Since the band gap energy is typically some large number of $k_B T$ even at room temperature in semiconductors and insulators, this result also justifies the use of the Maxwell-Boltzmann approximation for calculating the behavior of carriers in such pure or “intrinsic” materials.

C.7 Doping and Fermi levels

Donor atoms can be regarded as being something like a hydrogen atom in that they have one “extra” electron that overall is orbiting a net charge of $+e$ (the electronic charge magnitude). (That net charge is from all the protons in the atom’s nucleus, and all the other electrons in the donor atom.) Compared to the binding in a hydrogen atom, that binding of an electron to a donor atom is relatively weak; this weakness is the result of two pieces of physics. First, because semiconductor materials tend to have at least moderately large “relative dielectric constants ϵ_r ” (e.g., of the order of 10 or so), the Coulomb attraction force is weakened in proportion. Second, because the electron can behave as if it has an effective mass that is often significantly smaller than the free electron mass, that smaller effective mass tends to reduce the binding energy also, for example in some hydrogenic model of binding energy⁹. Such effective “binding

⁹ If we pretend we can treat the added electron “orbiting” the donor nucleus with a net single positive charge in the same way we modeled the hydrogen atom, but with a dielectric constant of $\epsilon_r \epsilon_0$ and with an effective mass m_{eff} , then our Rydberg formula would be

energies” can range from a few milli-electron-volts (meV) to a few 10’s of meV in typical semiconductors.

So, adding this donor atom adds an energy level that is just below the conduction band, by a few to a few 10’s of meV, and an electron that can either occupy this level or can be thermally excited into the conduction band. Whether or not it is thermally excited is determined by the Fermi level and Fermi-Dirac statistics.

Adding in such donors moves the Fermi level towards or even into the conduction band, which has to happen if there are electrons occupying either the donor level itself or the conduction band states¹⁰. So, a better picture of *n*-type doping is that we are adding in electrons and also energy levels, for electrons, just below the conduction band. If we analyze such a picture with the Fermi-Dirac distribution, we will explain the basic phenomena of doping, including being able to calculate just how many electrons are added to the conduction band (or “ionized” into the conduction band) at given doping densities and temperatures. We will not do that analysis here, though it is a straightforward extension of the models we have constructed.

For *p*-type dopants we end up with a similar model, but viewed “upside-down”. Such acceptor dopants add energy levels for holes just above the valence band, those holes may or may not be thermally excited into the valence band, and the addition of these *p* dopants moves the Fermi level nearer to or even into the valence band.

Incidentally, such an approach can also explain why adding impurities to wide band gap materials like insulators or wide-band-gap semiconductors does not always lead to conductivity. The key parameter is just where the dopant energy levels lie relative to the band edges. In such materials, those added energy levels are often quite far from the band edges, so the electrons or holes they add do not always get thermally excited into the corresponding band. Many such impurities only add what are called “deep levels” that are far from either band edge.

At low doping levels in semiconductors, often we can continue to use the “non-degenerate” Maxwell-Boltzmann approximation because the Fermi level can be quite far from the band edges¹¹. With high doping, we can however also get into the “degenerate” regime, as illustrated in Fig. C.2.

$$Ry_{eff} \approx \frac{m_{eff}}{2} \left(\frac{e^2}{4\pi\epsilon_r\epsilon_o\hbar} \right)^2 . \text{ With an effective mass } m_{eff} \text{ of, say, } 1/10 \text{ of the free electron mass,}$$

and a relative dielectric constant of, say $\epsilon_r \sim 10$ $\epsilon_r \sim 10$, then the effective Rydberg would be ~ 1000 times smaller than the hydrogen atom Rydberg, at more like 13.6 meV than 13.6 eV. This kind of hydrogenic approach to donors (and acceptors) is actually quite a good first model for such systems and their energy levels.

¹⁰ For example, at near zero temperature, the donor level would have to be “full” because all electron states below that are already filled, so at low temperatures the Fermi level would have to be essentially at the energy of the donor level or just above it.

¹¹ Even though the dopant energy levels are close to the band, and even though at low temperature the Fermi level would lie very close to the dopant energy level, at higher temperatures and low to moderate amounts of doping, the Fermi energy can lie quite far from the corresponding band edge, significantly “below” the doping energy levels. Incidentally, the

term “dopant level” could refer to the energy level of the dopant, or it could refer to the density of dopant atoms added to the material, though “doping level” would be a better term for this latter idea.

Appendix D

Quiz solutions

Chapter 2

Q2.2.1. Speed and energy are magnitudes without direction. So speed and both kinetic and potential energy are not vectors; these are all "scalar" quantities - that is, quantities that can each be represented by one number. Only velocity and momentum are vectors in this list, both having magnitude and direction.

Q2.2.2 Kinetic energy $= (1/2)mv^2 = (1/2) \times 60 \times (7000 \times 7000) = 1.47 \times 10^9 \text{ J}$

Q2.2.3 Momentum is mass times velocity, so the units are kg (for mass) times m/s (for velocity).

Q2.2.4

- a) For a gain in height of h , we have a gain in potential energy $= mgh = (60 \text{ kg})(10 \text{ m/s}^2)(2 \text{ m}) = 1200 \text{ J}$.
- b) Sorry if this is a bit of a trick question, but your *potential* energy gain is the same as if you went up slowly, since each time your height increased by the same amount, so it is still 1200 J.
- c) You have not gained any height, so you have not gained any potential energy, so the answer is 0 J.
- d) Note that the question is asking for the change in potential energy *starting from the top of the first set of stairs*, so h in mgh is 2 m, and the answer is $(60 \text{ kg})(10 \text{ m/s}^2)(2 \text{ m}) = 1200$
- e) The sofa does not move vertically, so there is no change in its gravitational potential energy. Of course, you have done a lot of work. That work is force times distance, so $200 \text{ N} \times 3 \text{ m} = 600 \text{ J}$, but that has just been dissipated as heat from friction.

Q2.2.5 There are at least two equivalent ways of attacking this problem.

(1) The "classic" (and much longer) method: For a height h and a mass m , the potential energy is $U = mgh$. The total magnitude of the gravitational force is $F = mg$. Then we see algebraically that $F = U/h$, where F is the total gravitational force acting on the ball. We can decompose that force into two parts at right angles to one another: one part is perpendicular to the plane (or "normal" to the plane) and the other part is along the tilted surface of the plane, so a surface at some angle to the horizontal (the angle of inclination of the plane). Let f be the

force down the tilted surface of the plane, and let L be the length of the plane. Then, resolving forces,

$$f = F \sin(\theta) = (U / h)(h / L) = U / L = (10 \text{ J}) / (4 \text{ m}) = 2.5 \text{ N}.$$

(2) The “smart” method: To get the ball to the top of the plane, with the known potential energy (increase) of $U = 10 \text{ J}$ in doing so, I must have pushed with some force f_{up} up the plane through the known distance $L = 4 \text{ m}$, so we must have $f_{\text{up}} L = U$, and hence $f_{\text{up}} = U / L = 10 / 4 = 2.5 \text{ N}$. That must be balancing the force pushing the ball down the plane, so the magnitude of force pushing the ball down the plane is $f = 2.5 \text{ N}$. (Note the “height” term that you might have thought you needed does not appear and neither do the angle of inclination of the plane and the acceleration due to gravity (g).)

Q2.4.1 Angular frequency $\equiv \omega = 2\pi f$ where f is the frequency, so $2\pi \times 100 \cong 628$ radians per second

Q2.4.2 The equation of motion for the simple harmonic oscillator is given as

$$\frac{d^2 y}{dt^2} = -\frac{K}{M} y = -\omega^2 y. \text{ So } \lambda = \omega^2 = (2\pi f)^2 = 4\pi^2 \times 10^2, f \text{ being the frequency of the oscillation in Hz.}$$

Q2.4.3 The definition of a linear operator A is that it is an operator for which

$A(\mathbf{x} + \mathbf{y}) = A\mathbf{x} + A\mathbf{y}$ and $A(c\mathbf{x}) = cA(\mathbf{x})$, where $\mathbf{x}$ and $\mathbf{y}$ are functions and c is a number. So,

- a) (Necessarily) True
- b) Not Necessarily True
- c) Not Necessarily True
- d) (Necessarily) True

Q2.5.1 Answer c. The wave velocity on a string under tension is $v_{10} = \sqrt{T / \rho}$. So increasing the tension by a factor of 9 from 10N to 90N increases the wave velocity by a factor of $\sqrt{9} = 3$.

Q2.5.2

- a) $-h(z - vt)$ – True. First, $-h(z - vt)$ is also just a function of $z - vt$, so it too must be a solution of the wave equation, since any function of $z - vt$ is a solution of the wave equation. Second, even if we thought that there were some restrictions (such as boundary conditions), on this wave equation (such as walls that the string is tied to) that therefore restrict possible solutions, because $h(z - vt)$ is a solution of the wave equation and the wave equation is linear, multiplying $h(z - vt)$ by any number, including in this case -1, gives a function that is also a solution.
- b) $h(z) - vh(t)$ – False. This kind of relation would apply if h was a linear operator on the “functions” z and t , in which case $h(z) - vh(t) \equiv h(z) - h(vt) \equiv h(z - vt)$, but here the h is a function, not a linear

operator. We have no reason to assert that this proposed relation should work as a solution to the wave equation, and generally it does not.

- c) $h(z + vt)$ – True. We know that any function of the form $g(z + vt)$ is a solution to the wave equation (it describes a wave propagating to the "left" – that is, to smaller positive z). Therefore, $h(z + vt)$ is also a solution of the wave equation.

Q2.6.1

- a) $\frac{d^2 y(z)}{dz^2} = -k^2 y(z)$ is a general statement of the Helmholtz equation
- b) $\frac{-1}{k^2} \frac{d^2 y(z)}{dz^2} = y(z)$ is a simple rearrangement of this, so it also is a version of the Helmholtz equation.
- c) $\frac{d^2 y(z)}{dt^2} + k^2 y(z) = 0$ has a derivative with respect to time not space (z), so it is not a version of the Helmholtz equation.

Q2.6.2 We construct the Helmholtz equation from the wave equation by assuming that the wave is sinusoidal with a particular angular frequency ω and mathematically we are free to choose any (real) value for that.

Q2.6.3 The relation between ω and k is $k^2 = \omega^2 / v^2$ (and, since conventionally we typically consider both ω and k to be positive numbers we also typically can write $k = \omega / v$). Therefore, if we double the ω , the corresponding k is also doubled.

Q2.7.1 The correct answer is (c). The boundary conditions $y(0) = 0$ and $y(L) = 0$ imply that k should have the form of $n\pi / L$, where n is integer. This is so we fit in integer numbers of half waves between the walls.

Q2.7.2 The correct answer is (c). The boundary condition $y(0) = 0$ implies that the wave function can only have the form $\propto \sin(kz)$. Since $\cos(kz)$ is 1 for $z = 0$, $\cos(kz)$ any amount of in the solution will mean that $y(0) \neq 0$, which violates the boundary condition there.

Q2.8.1

- a) True. For the oscillation of the lower frequency, the middle spring stays the same length at all times, so any force it exerts on the two masses does not change at all. In the problem, we are told to assume that, at rest in its equilibrium position, there is actually no compression or tension in any of the springs. In this (lower frequency) mode of oscillation, the two masses are moving equally in the same direction, so the length of this middle spring does not change at all, so it is exerting no force on either. So the situation is the same as as if there is no spring at all between the two masses. Therefore, the oscillation frequency of the system is the same as one mass attached to one wall with one spring. Incidentally, it makes no difference if there is actually tension or compression in the springs at the equilibrium position. We would simply be adding a constant and balanced set of forces on the masses, and would only

actually be concerned with the additional forces from changes in spring lengths since that would be all that accelerates the masses. If the springs are linear (and we assume they are) these further static force contributions do not matter.

- b) True. Just as in the answer to the preceding question, since the length of this spring does not change in this oscillation mode, it does not matter what its spring constant or "stiffness" is; it is exerting no (additional) force.
- c) False. The frequency difference between two mode is $2\pi\sqrt{3k/m} - 2\pi\sqrt{k/m}$. Therefore, doubling the mass changes the difference by a factor $1/\sqrt{2}$.

Q2.9.1

- a) The inner product is $1 \times 1 + (-1) \times 0 = 1 \neq 0$. Therefore, they are not orthogonal.
- b) The inner product is $2 \times 6 + (-3) \times 4 = 0$. Therefore, they are orthogonal.
- c) The inner product is: $1 \times 1 + (-1) \times (-1) + 2 \times (-1) = 0$. Therefore, they are orthogonal.
- d) The inner product is $1 \times 0 + 0 \times (-1) + 0 \times 0 = 0$. Therefore, they are orthogonal.

Q2.9.2

- a) The inner product is $2 \times 6 + (-3) \times 4 = 0$. Because the vectors are both non-zero and orthogonal to one another (so with an inner product of zero), they can form a complete orthogonal basis.
- b) These two vectors are, first of all, not orthogonal because their inner product is non-zero. Specifically $6 \times (-2) + 18 \times (-6) = -120 \neq 0$. So they cannot form an orthogonal basis for this reason alone. Now, it is possible for vectors to form a basis even if they are not orthogonal; we can still describe other vectors in terms of them. But because $x_1 = -3x_2$, these vectors are actually parallel to one another, so they cannot possibly describe any vector that is orthogonal to them, and hence for this second reason they cannot form a basis, even a non-orthogonal basis.

Q2.9.3

- a) The inner product is

$$\langle \mathbf{x}_1 | \mathbf{x}_2 \rangle \equiv [1 \quad -i]^* \begin{bmatrix} -1 \\ -i \end{bmatrix} \equiv [1 \quad i] \begin{bmatrix} -1 \\ -i \end{bmatrix} = -1 - (i)^2 = -1 - (-1) = -1 + 1 = 0$$

(See the text for a formal description of the Dirac notation $\langle \mathbf{x}_1 | \mathbf{x}_2 \rangle$, which gives us a convenient shorthand for writing down an inner product.)

- b) The inner product is

$$\langle \mathbf{x}_1 | \mathbf{x}_2 \rangle \equiv [1+i \quad -1]^* \begin{bmatrix} 1+i \\ 2 \end{bmatrix} \equiv [1-i \quad -1] \begin{bmatrix} 1+i \\ 2 \end{bmatrix} = (1-i)(1+i) - 2 = (1+1) - 2 = 0$$

Q2.9.4

- a) The Hermitian adjoint of this matrix is $H^\dagger = \begin{bmatrix} -i & i \\ -1 & 3 \end{bmatrix}$, which is not the same as the original matrix. So this matrix is not Hermitian.

- b) The Hermitian adjoint of this matrix is $H^\dagger = \begin{bmatrix} 1 & -i & 0 \\ -i & -8 & i \\ 0 & i & 0 \end{bmatrix}$, which is not the same as the original matrix, so this matrix is not Hermitian (and which illustrates that symmetry is not enough for a matrix to be Hermitian).
- c) The Hermitian adjoint of this matrix is the same as the matrix. (Note that $(1-i)^* = 1+i$.) Therefore, it is Hermitian.

Chapter 3

- Q3.1.1 True. By looking at the figure of black body radiation for 5800K in the text, we can see that power per unit wavelength at 500nm is much higher than at 1500nm. (The peak emission wavelength for this black body radiation is actually near 500 nm.
- Q3.1.2 False. By looking at the figure of black body radiation for 3000K in the text, we can see that power per unit wavelength at 500nm is lower than 700nm. (The wavelength of peak black body emission at 3000K is well into the infrared.)
- Q3.1.3 The correct answer is a. The relation between the energy (in joules) of a photon and the frequency f is $E = hf$.
- Q3.2.1 The formula for the energy of a hydrogen atom state (relative to the energy at which the hydrogen atom would just be ionized) is given by $E_H(n) = -\frac{Ry}{n^2}$ where $Ry \cong 13.6\text{eV}$ is the Rydberg energy, so for $n = 1$ the energy is -13.6 eV .
- Q3.2.2 $\Delta E = E_2 - E_1 = (-13.6\text{ eV})\left(\frac{1}{2^2} - \frac{1}{1^2}\right) \cong 10.2\text{ eV}$
- Q3.2.3 $E = hf = \frac{hc}{\lambda}$ in joules, if we take the usual definitions of Planck's constant h and the velocity of light c . To get to an energy in electron volts (eV) we have to divide by the magnitude of the electronic charge $\cong 1.602 \times 10^{-19}\text{ C}$. So, we have
- $$E \cong \frac{6.626 \times 10^{-34} \times 3 \times 10^8}{500 \times 10^{-9} \times 1.602 \times 10^{-19}} \cong 2.5\text{eV}$$
- Q3.2.4 Presuming the electron is a point particle in a circular orbit is not correct, and would actually lead to incorrect conclusions that simply do not agree with other phenomena we observe. For one thing, a “point” electron in such an orbit would have to radiate electromagnetic energy all the time because that motion would correspond to an accelerating charge. (Changing direction, as has to happen in a circular orbit all the time, corresponds to an acceleration.) When we probe atoms in various ways, we do not see any such orbits of a point particle, and later models, especially the Schrodinger equation, give an explanation in terms of wavefunctions and “fuzzy clouds” that apparently agrees very well with essentially

everything we do observe about atoms, but does not have any such actual orbits of this type.

Q3.2.5 $k = 2\pi / \lambda = 2\pi / (h / p) = 2\pi p / h$, so doubling the momentum p doubles the k value.

Q3.3.1

- a) False
- b) True
- c) False
- d) True

Q3.5.1 In each case, the question is whether the expression is equivalent to

$$E_n = \frac{\hbar^2}{2m} \left(\frac{n\pi}{L_z} \right)^2 \text{ for some positive integer } n$$

- a) True
- b) False
- c) False
- d) True

Q3.5.2 In each case, the question is whether the expression is equivalent to

$$\psi_n(z) = A_n \sin\left(\frac{n\pi z}{L_z}\right) \text{ for some positive integer } n \text{ and some constant } A_n. \text{ (The}$$

specific constants in the wavefunctions in this problem have all been chosen to be $\sqrt{2/L_z}$, which means that, if these are allowable wavefunctions, they are also normalized.)

- a) True. This corresponds to the wavefunction for $n = 1$
- b) False. This wavefunction would also not be zero at the walls.
- c) False. This wavefunction would correspond to $n = 3/2$, which is not an integer.
- d) False. This wavefunction would also not be zero at the walls.

Q3.5.3 To be a normalized solution, the wavefunction must be in the form

$$\sqrt{\frac{2}{L_z}} \sin\left(\frac{n\pi z}{L_z}\right) \text{ where } n \text{ can be any positive (and non-zero) integer. So,}$$

- a) False.
- b) True (for $n = 4$)
- c) False.
- d) False.

Q3.6.1

- a) The diffraction angle increases. Because the relation between wavelength λ and diffraction angle θ_c is $\theta_c \sim \lambda / 2\pi d$, θ_c increases when λ increases.

Generally, diffraction effects are stronger for longer wavelengths - intuitively, the wavelength is becoming larger compared to the physical dimensions of some object (such as a slit), so the wave-like effects are becoming relatively stronger.

- b) Because the relation between width d and diffraction angle θ_c is $\theta_c \sim \lambda / 2\pi d$, θ_c decreases when d increases.
- c) The relation between wavevector and wavelength is $\lambda = 2\pi / k$. Therefore, λ decreases when k increases, which means that θ_c decreases.
- d) Using the relation $\theta_c \sim \lambda / 2\pi d$, for $\lambda = 1\mu\text{m}$ and $d = 3\mu\text{m}$, we have $\theta_c \sim 3^\circ$.
- e) Due to de Broglie's relation $p = \hbar k = 2\pi\hbar / \lambda$, if we increase the momentum we reduce the wavelength, and as a result we are reducing the θ_c .
- f) According to uncertainty principle ($\Delta p_x \Delta x \geq \hbar / 2$) if we are increasing d here, we are decreasing Δp_x . Also, intuitively, the wavelength is becoming smaller compared to the dimension of the slit (because we are increasing the size of the slit), so wave effects are becoming less strong, so the diffraction angle is getting smaller. The diffraction angle is a measure of the uncertainty of the momentum in the "transverse" (here, x) direction.

Q3.6.2 The "intensity" or brightness fringes in the Young's slits experiment are spaced by $d_s = \lambda z_o / s$. For $s = 100\mu\text{m}$, $z_o = 30\text{cm}$, and $d_s = 15\mu\text{m}$, our wavelength is therefore 5 nm.

Q3.6.3 For an electron

- a) Mass is an intrinsic attribute of the electron
- b) Momentum is an attribute of the state of the electron
- c) Charge is an intrinsic attribute of the electron
- d) Whatever "shape" we can discuss for an electron, it is certainly an attribute of the state. Generally, shape-like properties are given by the wavefunction, as in the case of the "particle in a box" wavefunctions, or the orbitals of the hydrogen atom. Even then, shape is not as clearly defined as in normal classical "particles", like a brick for example. With quantum mechanical states generally we can only measure probability distributions for a particle like an electron, and even then those may not have definite boundaries, as in the case of hydrogen atom orbitals, for example, which formally extend arbitrarily in space even if they become very weak for large distances from the proton.
- e) Again, location, to the extent that it is even very definite, is given through the wavefunction or the corresponding probability density, which is an attribute of the state. Even then, position will not generally be very definite because the probability distribution will spread in space to some degree.

Q3.6.4 For a photon

- a) We would need to know the frequency f of the photon to know its energy $E = hf$. That is an attribute of the state (which we often call a "mode" for the case of a photon) that the photon is occupying.
- b) All photons have zero charge. The fact that photons have zero charge is different from saying it is impossible to assign a charge value to the photon. We can indeed assign a charge to the photon, and that charge is independent of what state it is in, so it is an intrinsic property of the photon, even if its value is always zero. (Sorry if this seems like a trick question!)

- c) Again, just like for energy, the wavelength of the photon is an attribute of the state (or "mode") it is in.
- d) Again, just like for energy or wavelength, the momentum of the photon is an attribute of the state (or "mode") that the photon is in.
- e) See the comment under "photon charge" question (part (b) above). All photons quite definitely have a mass of zero, so mass is an intrinsic property of the photon (albeit with a zero value). (Again, sorry if this seems like a trick question!)

Q3.7.1

- a) True. The wavefunction generally penetrates further into the barrier as the energy increases up towards the top of the barrier.
- b) True. The "decay constant" κ becomes a smaller number as the energy increases up towards the top of the barrier. That means the wavefunction is penetrating further into the barrier. Note that the wavefunction decay length (i.e., the distance z by which the wavefunction has decreased by a factor $1/e$) is formally given by $1/\kappa$.
- c) False. Because the barrier is infinitely thick, the magnitude of the reflection coefficient of the wavefunction is 1. The wavefunction cannot penetrate all the way through an infinitely thick barrier, so it is in the end always totally reflected. Because the magnitude of reflection coefficient is 1, we do get points of complete destructive interference on the "left", between the incident and reflected waves, and at those points, the wavefunction amplitude, and, therefore, the probability density, are indeed zero. Generally, students often find this paradoxical at first. How can the electron be 100% reflected from the barrier if it "spends some time" in the barrier? However, even if the electron spends some time inside the barrier, it always comes back out again, so the probability that it is reflected is 1. Actually, we cannot discuss ideas of "spending some time" inside the barrier when we are discussing monochromatic waves, because those go on forever in space and time. To have a meaningful discussion of this idea, we would have to consider pulses, which are not monochromatic. We can do that in quantum mechanics, and it is not too hard to do so, but we would need to go beyond what are considering here, which is only Schrodinger's "time-independent" wave equation. We would need his "time-dependent" wave equation.
- d) False. From the formula $\kappa = \sqrt{2m(V_o - E)}/\hbar$, κ does not go to infinity as E goes to zero, and it would have to go to infinity if there was to be no penetration of the wave into the barrier so that we would have zero penetration depth. To have κ go to infinity, the barrier would need to be infinitely high.

Q3.7.2

- a) Increasing the thickness of the barrier reduces the probability that the particle is transmitted, so it must increase the probability that it is reflected.
- b) As the barrier width increases, the magnitude of the reflectivity increases, which means that the amplitude of the incident and reflected waves are becoming closer in magnitude. That increases interference effects, both increasing the maxima and pushing the minima down towards zeros. If we had a reflectivity of magnitude 1, then at points on the "left" of the barrier, we

would actually have zeros in the wavefunction and probability density. So, paradoxically, increasing the magnitude of the reflectivity by increasing the barrier width actually leads to even "lower" low points at these interference minima.

- c) See the answer to part (b) above.
- d) On the right of the barrier, the particle's probability density is constant (a flat line) for a given barrier width (there is no wave coming from the right to lead to interference). The height of this line (i.e., the probability of finding the particle to the right of the barrier) decreases as the barrier thickens because it gets "harder" for the particle to tunnel through the thicker barrier.
- e) Unlike the situation on the left of the barrier, as the barrier width increases and tunneling through the barrier gets harder, the probability density right of the barrier decreases everywhere (since the distribution right of the barrier is constant for a given barrier width).

Chapter 4

Q4.2.1 Choice (a). The electron *very* much lighter than the proton.

Q4.2.2 Choice (d). Quite generally in such classical problems, the pair of masses orbits around their center of mass, which is like the position of the "fulcrum" (pivot point) if you were to put one particle on one end of a lever (or a children's see-saw) and the other particle on the other end, adjusting the fulcrum until they balanced (presuming the lever has no mass, of course). Since the proton is so much more massive than the electron, this center of mass is very close to the proton position. Incidentally, if we asked the same question for an electron and a positron (the anti-particle of the electron, with a positive charge), the center of mass would be exactly in the middle between them. Such an electron-positron pair could then form a "positronium" atom, rather like the hydrogen atom.

Q4.2.3 Choice (a). The reduced mass μ for two masses m_1 and m_2 is $\mu = \frac{m_1 m_2}{m_1 + m_2}$

or, equivalently (and more intuitively), $\frac{1}{\mu} = \frac{1}{m_1} + \frac{1}{m_2}$. If we make one of these

masses, say m_2 , larger, $1/\mu$ becomes smaller, and so μ becomes larger as we increase the size of one of the masses.

Q4.3.1 Choice (b). For the hydrogen atom, $V(r) \propto -1/r$, so decreasing r makes $V(r)$ decrease (in the sense we are using here), becoming a larger negative number.

Q4.3.2

- a) True. The polar angle θ decreases from 90 degrees at the equator to 0 degrees at the North Pole.
- b) False. If you are walking due north (or south), your azimuthal angle is not changing.

Q4.3.3

- a) Your polar angle θ is not changing if you walk due east (or west). You are just staying in a plane that is parallel to the equatorial plane (the x - y plane in the way we are showing spherical polar coordinates).
- b) The azimuthal angle is ϕ , and indeed it increases as we move East from the x - z plane.

Q4.4.1

- a) True. By adding 2π to ϕ , we have then simply gone round in a circle around the polar axis and are back where we started.
- b) True. Adding 2π to θ is like walking (and/or swimming!) round the earth along a line of longitude until we get back to where we started. (Note that, conventionally, we typically only use a range from 0 to π for θ when describing positions on a sphere, of course, because we do not need any larger range than that to define a position.)

Q4.4.2 Note quite generally that we require $-l \leq m \leq l$.

- a) Allowed
- b) Not allowed (the magnitude of m cannot be greater than l)
- c) Allowed
- d) Allowed
- e) Not allowed (m is too large)

Q4.4.3

- a) l corresponds to the total number of nodal circles in the spherical harmonic, so $l = 4$
- b) m is the number of nodal circles passing through the poles, so $m = 3$
- c) With a total of 4 nodal circles, three of which pass through the poles, then one does not pass through the poles. By symmetry, that last nodal circle has to lie round the equator, so the answer is one.

Q4.4.4

- a) $l = 0$
- b) $l = 3$
- c) $l = 2$
- d) $l = 1$
- e) $l = 0$

Q4.5.1

- a) For the energy of the Hydrogen atom, we have $E_n = -13.6(\text{eV})/n^2$. For the 3s orbital n is 3. Therefore, $E_3 = -13.6(\text{eV})/9$.
- b) The energy of an orbital does not depend on l . (Note, incidentally, that the hydrogen atom is the only atom for which this is reliably true. For other atoms, generally the values of l and m can influence the eigenenergies of the orbitals.) Therefore, the energy of the 3s is the same as 3p, and is $E_3 = -13.6(\text{eV})/9$.

Q4.5.2

- a) By looking at the function of $R_{3,0}(\rho)$, we find that there are 2 such zeros, and we can also use the expression that the radial functions have $n - l - 1$ such zeros, with $l = 0$.

- b) By looking at the function of $R_{2,0}$, we find that there is 1 such zero, and we can also use the expression that the radial functions have $n - l - 1$ such zeros, with $l = 0$.
- c) By looking at the function of $R_{3,2}(\rho)$, we see that there are no such zeros, and we can also use the expression that the radial functions have $n - l - 1$ such zeros, with $l = 2$.

Q4.5.3

- a) False. n cannot be zero.
- b) False. l cannot be negative
- c) False. l cannot be greater than or equal to n . Equivalently we must have $l \leq n - 1$
- d) False. $|m|$ cannot be greater than l . e) This is a valid set of quantum numbers.

Q4.6.1

- a) False. n cannot be zero.
- b) False. l cannot be negative.
- c) False. l cannot be greater than or equal to n . Equivalently we must have $l \leq n - 1$
- d) False. $|m|$ cannot be greater than l . e) This is a valid set of quantum numbers.

Q4.8.1

- a) True.
- b) True.
- c) True.
- d) False. If this were to obey Madelung's rule, it would have to first complete the filling of the $4s$ orbitals (and there are 2 of those) before moving to the $3d$ orbitals. That is, the configuration would have to be $1s^2 2s^2 2p^6 3s^2 3p^6 4s^2 3d^4 \equiv [\text{Ar}] 4s^2 3d^4$

Q4.9.1

- a) The electron is a fermion, and fermions obey Pauli exclusion.
- b) The proton is a fermion, and fermions obey Pauli exclusion.
- c) The photon is a boson, and bosons do not obey Pauli exclusion.
- d) The neutron is a fermion, and fermions obey Pauli exclusion.
- e) The helium-4 nucleus is a boson, and bosons do not obey Pauli exclusion.

Q4.9.2 Since the marbles are distinguishable, there are 8 ways to split them up between the containers: we have 2 choices for the red's container, 2 for the green, and 2 for the blue, and $2 \times 2 \times 2 = 8$.

Q4.9.3 Once in the bank accounts, we don't distinguish between the different dollars. It doesn't matter which dollar bill you "put" in which bank account. Once you give the bank your dollar bill, the bank adds "1" to the register for the appropriate account. After that, the bank puts the dollar bill in a drawer or safe, or gives the dollar bill to another customer, or sends it back to the Federal government, but that makes no difference to your accounts; any "link" between the dollar bill and the value in your account is broken. So only this dollar value "in" each account now matters. So there are 4 ways of distributing 3 (whole) dollars between the two

accounts: 1) put all \$3 in checking; 2) put \$2 in checking and \$1 in savings; 3) put \$1 in checking and \$2 in savings; 4) put all \$3 in savings.

Q4.10.1 The correct choice is (b). The combined system has two eigenenergies. The one related to the “symmetric” mode has eigenenergy lower than E_{01} , and the one related to the “antisymmetric” mode has eigenenergy higher than E_{01} .

Q4.10.2

- a) True. As we can see when we compare the plotted wavefunctions, when comparing the lowest two energy eigenstates of this coupled system, in the lower energy one, the electron is more likely to be found in the barrier region than it is in the upper energy one. In fact, in the upper one, the wavefunction has to go through zero in the barrier, which makes it difficult for the wavefunction on the average to have a large magnitude inside the barrier.
- b) True. As we learned from the lecture, there is one symmetric solution and one anti-symmetric solution for the eigenfunctions.
- c) False. As the barrier width shrinks, the symmetries of the wavefunctions are retained.
- d) False. The “bonding” state is the symmetric mode, and it has lower energy. The fact that it has lower energy is consistent with it overall being less curved. Indeed it is precisely because the “bonding” state has lower energy that this kind of behavior corresponds to bonding. The systems have lower energy overall if they “bond” - that is, go into this particular symmetric coupled states. So bonding is energetically favorable, and it would take energy to “unbind” the coupled system.

Q4.11.1

- a) True. We see that these centers are indeed regularly spaced in all three dimensions, and hence are a crystal lattice (of points) in the sense we are using here.
- b) False. The spheres (the “oranges”) do not fill all space – there is space between them (other than for the points at which they touch). Hence the spheres are not unit cells for this lattice.
- c) True. According to definition of the crystal, the crystal is periodic, and we are not able to notice any difference if we move by a period of the lattice.

Q4.11.2

- a) False. The diamond lattice is a set of two interlocking face-centered cubic lattices. In each of those cubic lattices there are some atoms that lie on the face of the cube and some that lie on the corners of the cubes. So it is not the case that the atoms lie on the corners of non-overlapping cubes.
- b) True. We can see from the picture of the diamond lattice that we could build the entire lattice by stacking the cubes whose outlines are shown in the picture, and those cubes would fill all space (there would be no space between the cubes).

Q4.11.3 True. Because we should not be able to tell any difference if we displace the crystal by one “spacing” between the lattice points. Having said that, it is quite possible to have multiple different atoms in a crystal structure. However, in that case an identical group of (possibly different) atoms has to be associated with each (mathematical) lattice point.

Q4.11.4 True. Carbon can be in a kind of hexagonal lattice (in graphite) or in a diamond lattice. Incidentally, these are both three dimensional lattices. There is also much interest in “graphene” sheets, which have a two-dimensional kind of hexagonal lattice. When those sheets are stacked on top of one another in a particular way, they form the graphite lattice.

Q4.11.5 True (by definition).

Q4.11.6 False. A ternary alloy is made of three different kinds of atoms.

Q4.11.7 False. Metals for conductors and oxides for insulators often do not need precise crystalline forms to operate successfully.

Q4.13.1

- a) False. The problem here is the 2π . It certainly should not be there if this statement is to be necessarily true.
- b) False. This would require that the crystal was symmetric about $z = 0$. While a crystal can be symmetric in this way, there is no requirement that it has to be, so this statement is not necessarily true.
- c) True. The underlying requirement for a periodic potential with a "repeat length" a is that $V(z) = V(z + a)$. By induction, we can extend this, one period at a time, to prove that $V(z) = V(z + Na)$ for any integer N .

Q4.13.2

- a) In a crystal, the potential is periodic with period a , so this statement is necessarily true.
- b) The square of the modulus of the wavefunction, $|\psi(z)|^2$, is a physically measurable quantity and so should be the same in any unit cell. However, the wavefunction itself is not necessarily periodic, so this statement is not necessarily true.
- c) Since the "crystal" here is periodic as required, any measurable property should also be periodic, and $|\psi(z)|^2$ is the probability density, which is measurable, so this statement is necessarily true.
- d) $\frac{2\pi}{3a} = \frac{2\pi \times 5}{15a}$, and such a value of k is one of the allowed values of the form $\frac{2\pi n}{15a}$ for any (positive, negative, or zero) integer n and the chain length $15a$, so this statement is necessarily true.
- e) Allowed values have to be of the form $\frac{2\pi n}{15a}$ for n being an integer. n here would have to be 3.75, so this statement is not true.
- f) $n = 0$ in the expression for allowed values of k , i.e., $\frac{2\pi n}{15a}$, is an allowed solution, and corresponds to $k = 0$, so this statement is necessarily true.

Q4.13.3

- a) $\psi(z) = u(z)\exp(ikz)$ – True. This is the statement of Bloch's theorem.

- b) $\psi(z+a) = \exp(ika)\psi(z)$ – True. Indeed, this is often used as an alternate (and equivalent) form for Bloch's theorem.
- c) $\psi\left(\frac{z}{2}+a\right) = \exp\left(i\frac{k}{2}a\right)\psi\left(\frac{z}{2}\right)$ – False. Just because k is an allowed value, it is not necessarily the case that $\frac{k}{2}$ is an allowed value.
- d) $\psi(z+a) = u(z+a)\exp(ik(z+a))$ – True. This is easily proved by substituting $z+a$ into Bloch's theorem.
- Q4.14.1 False. For each choice of N atoms in a (e.g., one-dimensional) crystal, the allowed k values are equally spaced by $2\pi / Na$. Therefore, a larger crystal (so, with a larger N) will have a smaller spacing of the allowed k values.
- Q4.14.2 True. Because k values are periodic over a range of $2\pi / a$ with spacing of $2\pi / Na$, the maximum number of different k is equal to N , the number of unit cells (where we remember that, formally, we need to leave off the value at one or other extreme of the range). (Note that technically we are only counting the allowed k values in one Brillouin zone, which is the standard approach unless we specifically choose otherwise.)
- Q4.14.3 False. For each value of k , there will generally be multiple possible energy eigenstates, corresponding to states in the different bands in the crystal.
- Q4.14.4 True. We can extend the band diagram by moving curves “sideways” by $2\pi / a$ creating an “extended” zone scheme that simply repeats itself.
- Q4.14.5 False. In the band throughout the Brillouin zone we can fit different parabolic curvatures at different $\mathbf{k}$ and each such parabolic curvature would generally correspond to a different effective mass. Indeed, if there was just one effective mass throughout the Brillouin zone, the band would just be a parabola, which in general bands are not. Effective masses that are approximately constant over some small or moderate range of the Brillouin zone are useful, but those ranges do not extend through the entire Brillouin zone.
- Q4.14.6 True. For example, the approximately parabolic curvature of the valance band in many materials corresponds formally to a negative mass. Even within one band, the band will typically both curve “up” (a positive effective mass) and “down” (a negative effective mass) as we move through the Brillouin zone. So, indeed, the electron can have negative effective mass.
- Q4.14.7 True. The crystal momentum is just given by $p = \hbar k$
- Q4.15.1 False. The Bloch theorem states that we can write down the wave function in the form $u(\mathbf{r})\exp(i\mathbf{k}\cdot\mathbf{r})$ where $u(\mathbf{r})$ is unit cell function. The unit cell function is not in general just some constant A , but varies inside the unit cell.
- Q4.15.2 False. The Brillouin zone of a face-centered cubic lattice is not in fact cubic, as shown in Fig. 4.20.
- Q4.15.3 False. This is not necessarily the case. The silicon (lowest) conduction band, for example, has 6 equivalent minima.

- Q4.15.4 False. This is not necessarily the case. The minima in the silicon conduction band are neither at the center nor the edge of the Brillouin zone, for example.
- Q4.15.5 True. This is a characteristic of any material with what is called a “direct gap”. There are many such materials, including GaAs.
- Q4.16.1 False. A full band does not conduct electricity. The electrons cannot change states within the band because all the states are full.
- Q4.16.2 False. Semiconductors have smaller band gap energies as compared to insulators, and as a consequence they have non-negligible populations of electrons (and/or holes) that are thermally excited “across” the band gap. Therefore, at finite temperatures, even pure semiconductors conduct electricity to some degree.
- Q4.16.3 False. Such doping can reduce the number of electrons in the (otherwise essentially full) valence band, which frees up some states that then allow some conduction because electrons now can change states in that valence band. We can also view the same process in terms of holes, and in that picture we can view this as adding holes that can now conduct as if they were positive charge carriers. So, *p*-doping can indeed increase the electrical conductivity.
- Q4.16.4 False. Metals conduct electricity because their conduction band is partially filled. Therefore, electrons can change their state and participate in conduction. In metals, this is true even if the metal is completely pure.
- Q4.16.5 True. Arsenic is a Group V element, and effectively adds an electron when added to a Group IV semiconductor.
- Q4.16.6 True. Phosphorous is also a Group V element.

Chapter 5

- Q5.2.1
- a) Because each time we toss the coin there are two possible outcomes (heads or tails), the total number of microstates is $2 \times 2 = 4$.
 - b) The macrostate with 1 head and 1 tail has the maximum number of microstates, which is 2 (head-tail and tail-head).
- Q5.2.2 Because each time we toss the coin there are two possible outcomes (heads or tails), the total number of microstates is $2 \times 2 \times 2 \times 2 = 32$
- Q5.2.3 Because each time we toss the coin there are two possible outcomes (heads or tails), the total number of microstates is $2^8 = 256$
- Q5.2.4
- a) The total number of such microstates is: $\frac{8!}{4!4!} = 70$
 - b) The total number of microstates is: $\frac{8!}{2!6!} = 28$

- c) The most probable macrostate occurs at zero excess, which means that we have equal number of heads and tails (four of each). Therefore, the total number of microstates is $\frac{8!}{4!4!} = 70$

Q5.2.5

- a) For the most probable macrostate, using Stirling's approximation,

$$\frac{8!}{4!4!} \cong \frac{\sqrt{2\pi \times 8} \left(\frac{8}{e}\right)^8}{\sqrt{2\pi \times 4} \left(\frac{4}{e}\right)^4 \times \sqrt{2\pi \times 4} \left(\frac{4}{e}\right)^4} = 72.22$$

- b) Using Stirling's approximation, the total number of microstates is

$$\frac{8!}{5!3!} \cong \frac{\sqrt{2\pi \times 8} \left(\frac{8}{e}\right)^8}{\sqrt{2\pi \times 3} \left(\frac{3}{e}\right)^3 \times \sqrt{2\pi \times 5} \left(\frac{5}{e}\right)^5} = 57.93$$

Q5.2.6 For $1/e^2$ reduction of multiplicity, in this approximation, the number

$s_e = \sqrt{N}$, so we have $s_e = 8$. So, $2s_e = 16$.

Q5.3.1 This is True, because of the minus sign in the total energy $U(s_e) = -2\mu_s B s_e$.

Q5.3.2 False. The multiplicity of the entire system's macrostate is the *product* of the multiplicities of the subsystems' macrostates.

Q5.3.3 The correct choice is (b). The total system is insulated from its environment, so there is no way for energy to leave. (Energy might, however, flow from one (sub)system to another *within* the overall combined system.)

Q5.3.4

- a) True. The energy of the overall system is conserved (since it is insulated from its environment). The overall energy only depends on the spin excess, so the spin excess must remain the same.
- b) False. If the average energy per spin settled to zero, the total energy of the system would settle to zero. Since the overall system does not necessarily start off with zero energy, this is not necessarily the case.
- c) True. The average energy per spin does indeed settle down to be the same on both sides, so the fraction of spin-up particles on each side settles to the same number on the average, at least as far as that is possible given that we are taking ratios of integer numbers.

Q5.4.1

- a) False. There is no requirement that the multiplicities are equal in the two systems.
- b) True. This is proved above and is ultimately equivalent to the statement that systems in thermal equilibrium are at the same temperature.
- c) False. There is no such general requirement. Indeed, if the multiplicities are different, this will not be the case because it would contradict the correct result

$\left. \frac{1}{g_1} \frac{\partial g_1}{\partial U_1} \right|_{N_1} = \left. \frac{1}{g_2} \frac{\partial g_2}{\partial U_2} \right|_{N_2}$ in its magnitude, and the sign is also wrong compared to this correct result.

- d) True. This is mathematically equivalent to $\left. \frac{1}{g_1} \frac{\partial g_1}{\partial U_1} \right|_{N_1} = \left. \frac{1}{g_2} \frac{\partial g_2}{\partial U_2} \right|_{N_2}$, and is

effectively another version of the statement that the subsystems have the same temperature.

Q5.4.2 False. The combined system at equilibrium has higher entropy due to the second law of thermodynamics. The flow of heat energy from the hotter system to the colder system will have increased the entropy.

Q5.4.3 True. We remember that the rate of change of entropy with energy at a constant number of particles is exactly the partial derivative that defines

"1/temperature". Specifically, we have, for the fundamental temperature $\tau = k_B T$,

$$\frac{1}{\tau} = \left(\frac{\partial \sigma}{\partial U} \right)_{N}. \text{ So if this partial derivative is increased, the temperature has reduced.}$$

Q5.4.4 This is True, because in fundamental units, it is just the log of the multiplicity (which is just a number).

Q5.4.5 True. The multiplicity of the combined system is just the product of the multiplicities of the individual system. Given that the entropy of each system is just the logarithm of the multiplicity, and the logarithm of a product of two numbers is the sum of the logarithms of the individual numbers, then the total entropy here is just the sum of the two entropies.

Q5.4.6 False. The entropy of the system is the logarithm of the multiplicity of a system, and conversely, the multiplicity is the exponential of the entropy.

Q5.4.7

- There are two possible microstates: (i) system 1 with spin up and system 2 with spin down, (ii) system 1 with spin down and system 2 with spin up. Energy conservation does not allow any other possibilities when in a magnetic field.
- The most probable macrostate is when each of the system has one spin up and one spin down. We calculated the number of microstates for this macrostate in the preceding part of this question, counting 2 such microstates.
- Because the only possible microstates are one spin up in system 1 and one spin down in system 2 or one spin down in system 1 and one spin up in system 2, there are only 2 microstates, both of which are in the same macrostate. Therefore, 100% of the possible microstates of this system are in the most probable macrostate.

Q5.4.8 The entropy in fundamental units is $\log(100) \approx 4.6$. For interest, in

$$\text{thermodynamic units, this would be } k_B \log(100) \approx 6.4 \times 10^{-23} \left(\frac{J}{K} \right).$$

Q5.4.9

a) $\Delta S_1 = -\frac{1}{303.15} = -3.3 \times 10^{-3} \left(\frac{J}{K} \right)$. Note that energy is flowing *out* of body 1, so its entropy is decreasing.

b) Note energy is flowing *into* body 2, so its entropy is increasing.

c) $\Delta S = \Delta S_1 + \Delta S_2 = 0.1 \times 10^{-3} \left(\frac{J}{K} \right)$

Q5.4.10 $\Delta g = \exp(\Delta \sigma) = \exp(100) = 10^{(\log_{10} e) \times 100} \approx 10^{43.43}$ Q5.4.11 $\Delta g = \exp\left(\frac{\Delta S}{k_B}\right) = \exp\left(\frac{1}{1.38 \times 10^{-23}}\right) = 10^{\log_{10} e \times 7246 \times 10^{23}} \approx 10^{3.2 \times 10^{22}}$

Q5.5.1 False. Heat flowing out of the hot reservoir corresponds to a decrease of entropy of the hot reservoir.

Q5.5.2 True. The engine can be more efficient if the hot reservoir is hotter. We can see that from the formula for the efficiency $\eta_{engine} \leq 1 - \frac{T_C}{T_H}$. The entropy reduction

for a given transfer of heat energy out of the hot reservoir is smaller if the hot reservoir is hotter. That means that we need to transfer a smaller amount of heat into the cold reservoir just to cancel that entropy reduction. As a result, there is a larger fraction of the heat energy out of the hot reservoir that can be utilized for work without violating either the First Law of Thermodynamics (which requires conservation of energy overall) or the Second Law of Thermodynamics (which requires that the entropy does not reduce)

Q5.5.3 False. If we increase the temperature of the cold reservoir, then the limit to the efficiency becomes a smaller number. Again, we can see that from the formula for the efficiency $\eta_{engine} \leq 1 - \frac{T_C}{T_H}$. If the cold reservoir temperature is increased,

then we need a larger transfer of heat into the cold reservoir to achieve the necessary increase in its entropy to balance the decrease of entropy of the hot reservoir. That means that less energy is available for work.

Q5.6.1

a) $\frac{P_2}{P_1} = \exp\left(-\frac{\varepsilon_2 - \varepsilon_1}{k_B T}\right) = \exp\left(-\frac{25}{25}\right) = \exp(-1) = \frac{1}{e}$

b) $\frac{P_3}{P_2} = \exp\left(-\frac{\varepsilon_3 - \varepsilon_2}{k_B T}\right) = \exp\left(-\frac{50}{25}\right) = \exp(-2) = \frac{1}{e^2}$

c) $\frac{g_1}{g_2} = \frac{P_1}{P_2} = e$

Q5.7.1

a) False. At diffusive equilibrium, we presume maximum entropy so small transfers result in vanishingly small entropy changes

- b) False. At diffusive equilibrium, we presume maximum entropy so small transfers result in vanishingly small entropy changes
- c) True. Because the net change of the entropy is zero, if the entropy of one box increases the other decreases.
- d) True. The overall entropy of the system remains approximately constant because the net change of entropy is zero, or at least vanishingly small.

Q5.7.2 The answer is (c) $1/e$. The Gibbs factor corresponding to P_A is

$$\exp\left(\frac{1 \times 25 \text{ meV} - 0}{25 \text{ meV}}\right) = e. \text{ The Gibbs factor corresponding to } P_B \text{ is}$$

$$\exp\left(\frac{2 \times 25 \text{ meV} - 2 \times 25 \text{ meV}}{25 \text{ meV}}\right) = 1$$

Q5.8.1

a) True. The occupation probability is $\frac{1}{\exp\left(\frac{50-50}{25}\right) + 1} = \frac{1}{2}$

- b) False. The occupation probability of the electrons less than Fermi energy is larger than $\frac{1}{2}$.

c) False. The occupation probability is $\frac{1}{\exp\left(\frac{75-50}{25}\right) + 1} = \frac{1}{e+1} \approx 0.29$, which is

close to $1/3$ but not exactly $1/3$.

- d) True. The entire Fermi-Dirac distribution "steepens" as the temperature is reduced, increasing the occupation probabilities of all states below the Fermi level.

Q5.8.2

a) The correct answer is $\frac{1}{\exp\left(\frac{2}{0.025}\right) - 1} \approx \exp(-80) \approx 1.8 \times 10^{-35}$

b) The correct answer is $\frac{1}{\exp\left(\frac{2}{0.25}\right) - 1} \approx \exp(-8) \approx 3 \times 10^{-4}$

c) The correct answer is $\frac{1}{\exp\left(\frac{2}{0.5}\right) - 1} \approx \exp(-4) \approx 0.02$

d) The correct answer is $\frac{1}{\exp\left(\frac{100}{25}\right) - 1} \approx \exp(-4) \approx 0.02$

Q5.8.3 True. We can see this easily from the plotted functions because we are just shifting all three curves to the right.

Q5.8.4 False. For energies below the Fermi level, the occupation probabilities will decrease for the Fermi-Dirac distribution

Chapter 6

Q6.2.1 False. The electron *acceleration* is proportional to the applied field in ballistic transport.

Q6.2.2 True. The drift transport model, with its continual acceleration followed by a collision-induced reset to an average zero velocity, leads to an overall average velocity that is proportional to the applied field.

Q6.2.3 False. A shorter drift time (for the same mass) means that the electron gets less time to be accelerated by the field before its velocity is again reset to an average zero value. So its average velocity is less, so its mobility is lower.

Q6.2.4 True. For a given field, a lighter mass is accelerated faster so reaches a larger velocity before having the velocity reset to an effectively zero value. So smaller mass leads to higher average velocity for the same scattering time.

Q6.2.5 True. In drift transport, the average velocity is proportional to the field, which leads to a current that is proportional to voltage, which is Ohm's Law.

Q6.2.6 True. For the many "cancelling minus sign" reasons discussed in the text, holes in such a valence band maximum behave as if they had positive masses and positive charges, which leads to the same direction of electric current as for the electrons in the conduction band. (Note, though, that the electrons in the conduction band and the holes in the valence band are, in this case, going in opposite directions.)

Q6.4.1 True. There are many more holes in the *p*-doped semiconductor than there are in the *n*-doped semiconductor, so they will tend to diffuse into the *n*-doped semiconductor.

Q6.4.2 False. The acceptors are in the *p*-type material. (Also, though electrons and holes can move, the acceptor and donor atoms themselves are essentially fixed in position, at least after fabrication of the material, so they cannot in practice diffuse.)

Q6.4.3 True. The electric field that develops is in correct direction to try to push holes back into the *p*-type material.

Q6.4.4 False. It is a region with essentially no electrons or holes, but the acceptor and donor atoms remain exactly where they were. (It is possible to deliberately make a diode in such a way that most of the depletion region contains no acceptor or donor atoms, but that is called a *p-i-n* diode, with the "*i*" standing for "intrinsic". Such structures are often made for photodetectors because they give a thicker depletion region for absorbing photons.)

Q6.4.5 True. That is the thermal and diffusive equilibrium condition for the whole structure. (Actually, this is only formally correct in the dark when we are not shining light on the diode, but we have not yet discussed the effect of light, so this answer is correct at this point.)

- Q6.4.6 True. This is possibly the original definition by Gibbs. The full proof is given in Section 5.7 of the text.
- Q6.4.7 True. The reason for this is exactly because of the “Gibbs” definition of the chemical potential (Fermi level); by applying a voltage we are just changing the energy of the particles without changing any entropies.
- Q6.4.8 False. The electron diffusion is from the n-type to the p-type.
- Q6.4.9 True (by definition)
- Q6.4.10 True (by definition)
- Q6.4.11 False. Reverse leakage is caused by *minority* carriers diffusing.
- Q6.4.12 False. Though there is still some minority carrier diffusion “backwards” that slightly cancels out some of the forward current, the vast majority of the forward current is majority carrier diffusion.
- Q6.4.13 True. According to the ideal diode current-voltage formula,
- $$I = I_S \left[\exp\left(\frac{eV}{k_B T}\right) - 1 \right], \text{ for } \frac{eV}{k_B T} = \log 2, \text{ we then have}$$
- $$I = I_S [\exp(\log 2) - 1] = I_S [2 - 1] = I_S, \text{ so these magnitudes are equal.}$$

Chapter 7

- Q7.2.1 True. The voltage that just stops the current is the one that just stops all the electrons. The kinetic energy of the emitted electrons increases as the photon energy increases because there is more excess energy available (so the energy “above” the emission “barrier” or “work function” required to extract an electron from the metal. Therefore, increasing the photon energy increases the magnitude of the voltage required to stop all the electrons reaching the collecting plate.
- Q7.2.2 False. By increasing the power of the light, we do not increase the energy of the photons because the energy of the photons is just proportional to the light frequency not the power of the light. Hence the kinetic energy of the individual emitted electrons does not increase just because we increase the light power (at the same photon energy).
- Q7.3.1 True, because the separation between allowed k values in a given direction is inversely proportional to dimension in that direction.
- Q7.3.2 True, because it is proportional to the product of the density of states in each direction, and those are each proportional to the linear dimension in that direction.
- Q7.4.1 False. There is also the exponential factor in the denominator, which prevents the cubic increase with frequency from dominating at high frequencies, and this is actually what stops the “ultraviolet catastrophe”.
- Q7.4.2 True. This is the Stefan-Boltzmann law
- Q7.5.1 White, because it emits less radiation
- Q7.5.2 Black, because it absorbs more radiation.
- Q7.5.3 Black, because it emits more radiation

Q7.5.4 Neither one will change temperature (choice (c).) Though the emissivity of the white body will be less, so it emits less radiation to the black body than the black body emits to it, because the absorptivity of the white body is the same as the emissivity, so it absorbs less radiation from the black body. Hence, since the two are at the same temperature to start with, there is no change in the temperature of white body - its absorbed and emitted radiation are in balance. Also, the light not absorbed by the white body must be reflected, and given the terms of this question, there is nowhere for it to go other than back to the black body. So it also has its radiated and absorbed power in equilibrium. Anyway, quite generally, the second law of thermodynamics tells us that neither one of these two bodies can heat up the other since they start at the same temperature.

Q7.6.1

- a) False. There is still the possibility of spontaneous emission.
- b) True. Absorption and stimulated emission will balance out. It is indeed possible to make at least some materials more transparent by "pumping" them in this way.
- c) True. With $N_2 > N_1$, stimulated emission is more likely than absorption, so the outgoing light beam can have more photons than the incoming light beam.

Chapter 8

Q8.2.1 True. The carriers (electrons and holes) can keep moving in the field, giving current, until they recombine. Incidentally, that recombination can be in the "bulk" of the material itself, or it can be when the carriers reach the electrodes at the ends of the semiconductor that are applying the voltage. Generally, reaching the electrode effectively "recombines" the carriers.

Q8.2.2 False. Electrons and holes generated in the depletion region of the diode tend to move in the "built-in" field - that is the slope in the bands - "downhill" into one or other doped region (electrons move "downhill" towards the n-doped region, holes move "downhill" (in their terms) towards the p-doped region. That separation of charge itself builds up an electric field (which tends to oppose this movement of the charge) over the depletion region, and so there is a voltage change overall across the diode.

Q8.2.3 False. It is a "reverse" current. The generated electrons move downhill towards the *n*-doped region, which is in the opposite direction to the (forward) electron diffusion current in a diode. Similarly, holes move downhill (in their terms) towards the *p*-doped region, which is also in the opposite direction to the (forward) hole diffusion current.

Q8.2.4 False. Much or all of this photocurrent will be drift current from the large electric field in the depletion region of a reverse biased diode.

Q8.2.5 False. The photocurrent is in the same sense as the reverse leakage current, so it does not cancel it

Q8.2.6 True. Shining light on a photodiode tends to push it towards forward bias, but the photocurrent itself is in the direction of a "reverse" current.

- Q8.3.1 False. Heating up a material cannot invert the populations. Because of the Boltzmann factor, a higher energy level state always has a lower occupation probability than a lower energy state.
- Q8.3.2 False. It can still emit, both spontaneously and through stimulated emission. It is just that, for any given incident photon, it is more likely to be absorbed than to stimulate emission. Incidentally, the process of absorption followed by (usually spontaneous) emission is exactly what happens in fluorescent materials. Absorption raises the material from one level to another, and it then emits photons as part of the process of the material "falling" back into its original level.
- Q8.3.3 Direct gap. The emission process is generally much more efficient in direct gap materials
- Q8.3.4 False. Light-emitting diodes do not rely on being at high temperatures to emit light. (In fact, it is even possible to cool a material by light emission.)
- Q8.3.5 True. The main current in a forward-biased light emitting diode for both electrons and holes is diffusion current into the depletion region.
- Q8.3.6 False. Absorption of a photon can take the system from a low-energy state to a high-energy state.
- Q8.3.7 False. There is no such process, and if there were, it would not be clear where the energy would come from. The point of this question, actually, is to emphasize that there is indeed a kind of asymmetry here between absorption and emission processes. While stimulated emission and absorption are essentially exactly inverse processes, there is no corresponding "inverse" process for spontaneous emission.
- Q8.3.8 False. Lasers generally emit spontaneous emission as well, and in practice into many "modes". There is no easy way to turn that off. Quite generally, if we have finite probabilities of stimulated emission, we cannot avoid having finite probability of spontaneous emission because for the same mode the same basic constants are involved. The spontaneous emission into the same mode as the lasing is, however, actually essentially indistinguishable from the lasing emission - there is no "difference" between a spontaneous photon and a stimulated one into a given mode. We could try to suppress spontaneous emission into other modes, but we can't suppress it for the lasing mode without similarly suppressing the lasing itself. The larger problem with spontaneous emission in lasing is generally with spontaneous emission into other modes, but completely suppressing that is in practice very hard or essentially impossible.
- Q8.3.9 False. It may be hot, but need not be. Also, we cannot get inverted populations simply by heating something up.
- Q8.3.10 There has to be enough gain to overcome other losses in the system so that the net "round-trip" gain is greater than one, so the mere existence of some gain in gain medium is not enough. (Also, of course, we need some kind of resonator to feed the photons back into the gain medium to give the "round trip".)
- Q8.3.11 False!!

Appendix E

Solutions to selected problems

Chapter 2

2.2.1

The potential energy of a charge q in electrostatic potential U can be calculated as:

$$V = qU$$

The resulting force on the charge is given by:

$$F = -\nabla V$$

a) The calculated potential energy is $V = -(1x + 3y - 1.5z)$ J, and the force is $F = -(1\mathbf{i} + 3\mathbf{j} - 1.5\mathbf{k})$ N, where $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ represent unit vectors along x , y , z coordinates. The magnitude of the force is 3.5 N.

b) The first potential would generate force $F_1 = -2\mathbf{i}$ N with magnitude 2 N, while the second potential would generate force $F_1 = (-1\mathbf{i} - 1\mathbf{j})$ N with magnitude 1.41 N. So, the first potential imposes stronger force on the electron.

2.4.1

(i) Frequency (whether conventional frequency or angular frequency) has units of $(1/\text{s})$ where “s” means “seconds”. Given that $\omega = \sqrt{K/M}$ for such an oscillator, rearranging to solve for K gives: $K = \omega^2 \left(\frac{1}{\text{s}^2} \right) \times M(\text{kg})$. The SI units of K are therefore $\frac{\text{kg}}{\text{s}^2}$.

(ii) Given that a newton has units of $\frac{\text{kg} \times \text{m}}{\text{s}^2}$, deduced by remembering that Force (N) = mass (kg) $\times$ acceleration (m/s^2), the units of K can also be written as $\frac{\text{N}}{\text{m}}$.

2.4.5

For a mass on a spring motion of some peak amplitude and starting position y_m ,

$$y = y_m \cos(\omega t), \text{ where } \omega = \sqrt{K/m} = \sqrt{10/2} \approx 2.24$$

Note that cosine is used instead of sine here because the spring is at its maximum position at $t = 0$. In this case, $y = 1$ and $y_m = 2$, so $\cos(2.24t) = 1/2$ or $t \approx (1/2.24)\cos^{-1}(1/2) \approx 0.467$ s. So, it takes ~ 0.467 s for the mass to travel back to $y = 1$ m.

Note that in this question we did not even specify whether the mass is hanging from a spring in normal Earth's gravity, but it makes no difference if it is. At equilibrium, gravity would pull on the mass, of course, stretching the spring by some amount y_0 . With the acceleration due to gravity as g , equating the gravitational and spring forces in equilibrium gives $mg = Ky_0$. Thus, the spring is stretched by $y_0 = mg/K$ m. But this stretching makes no difference to the oscillation frequency of the mass; the gravitational force is just a constant that leads to an overall displacement of the mass, but makes no difference to the restoring force back to this equilibrium position since the gravitational force is the same independent of the vertical position in such a problem. For completeness, we will include the gravitational potential energy in our solution below, though it is not necessary to do so.

To calculate the speed of the mass at this point one can use the conservation of energy:

$$\begin{aligned} & \frac{1}{2} K \left(y_m + \frac{mg}{K} \right)^2 \text{ (initial elastic potential energy)} \\ &= \frac{1}{2} K \left(y + \frac{mg}{K} \right)^2 \text{ (new elastic potential energy)} \\ &+ mg \Delta y \text{ (gravitational potential energy change from change } \Delta y \text{ in height)} \\ &+ \frac{1}{2} mv^2 \text{ (kinetic energy for the mass at the height } y) \end{aligned}$$

Expanding the squares gives

$$\frac{1}{2} Ky_m^2 + y_m mg + \frac{1}{2K} (mg)^2 = \frac{1}{2} Ky^2 + ymg + \frac{1}{2K} (mg)^2 + mg(y_m - y) + \frac{1}{2} mv^2$$

$$\text{which becomes } \frac{1}{2} Ky_m^2 = \frac{1}{2} Ky^2 + \frac{1}{2} mv^2$$

where we see that, indeed, the gravitational terms have cancelled out. We only need to consider the elastic potential energy that is in addition to the potential energy at the equilibrium position.

Putting in the given values, we have $0.5 \times 10 \times 2^2 = 0.5 \times 10 \times 1^2 + 0.5 \times 2v^2$

Rearranging allows us to solve for v^2 , giving $v^2 = 15$, and hence $v \approx 3.87$ m/s.

2.7.2

(a) From the classical wave equation, we know

$$v = \sqrt{\frac{T}{\rho}} = \sqrt{\frac{80}{8 \times 10^{-4}}} \approx 316.2 \text{ m/s}$$

For the fundamental mode, $\lambda = 2L = 120 \text{ cm}$. Also, $v = \lambda f$. Thus,

$$f_0 = \frac{v}{\lambda} \approx \frac{316.2 \text{ m/s}}{1.2 \text{ m}} \approx 263.5 \text{ Hz}.$$

For the second harmonic, $f = 2f_0 \approx 527 \text{ Hz}$.

(b) $f \propto v \propto \sqrt{T}$. Thus, frequency gets lower when the tension is decreased.

2.7.4

The Helmholtz equation is given by $\frac{d^2 y}{dz^2} = -k^2 y$. For a uniform string of length L stretched between two rigid posts, the solution is of the form $y(z) = A \sin(kz)$

where $k = n\pi / L$, $n = 1, 2, 3, \dots$

The three eigenfunctions that correspond to the lowest frequencies are those that correspond to $n = 1, 2$, and 3 because $k = \omega / v$ for wave velocity v . Therefore, the spatial forms of these three eigenfunctions are

$$y_1(z) = A \sin(\pi z / L), y_2(z) = A \sin(2\pi z / L), y_3(z) = A \sin(3\pi z / L)$$

2.9.5

(i) As in Section 2.8, we can construct equations analogous to Eqs. (2.65) and (2.66)

$$-\omega^2 m_L x_1 = -2Kx_1 + Kx_2 \quad \text{and} \quad -\omega^2 m_R x_2 = -2Kx_2 + Kx_1$$

based on notation as in Fig. 2.6. Multiplying the first equation by $1/K$ and the second by $m_L / m_R K$ and moving terms from left to right and *vice versa*, we can rewrite these equations as

$$2x_1 - x_2 = \frac{\omega^2 m_L}{K} x_1 \quad \text{and} \quad -\frac{m_L}{m_R} x_1 + 2\frac{m_L}{m_R} x_2 = \frac{\omega^2 m_L}{K} x_2$$

Defining $\rho = m_L / m_R$ and $\lambda \equiv \omega^2 m_L / K$, we have

$$2x_1 - x_2 = \lambda x_1 \quad \text{and} \quad -\rho x_1 + 2\rho x_2 = \lambda x_2$$

In matrix form, this becomes $M \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ with $M = \begin{bmatrix} 2 & -1 \\ -\rho & 2\rho \end{bmatrix}$ and

$$\lambda = \frac{m_L \omega^2}{K}$$

as required.

(ii) This matrix is *not* Hermitian. The two off-diagonal terms, -1 and $-\rho$, are *not* complex conjugates of one another. (This does have consequences for the solutions of the problem, as we will see in part (vi).)

(iii) We proceed as usual for the solution for eigenvalues. We bring the right hand side terms of the matrix equation into the matrix, leaving 0 on the right, i.e.,

$$\begin{bmatrix} 2-\lambda & -1 \\ -\rho & 2\rho-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

and use the condition that the determinant must vanish for a solution, i.e.,

$$(2-\lambda)(2\rho-\lambda)-\rho=0, \text{ i.e., } \lambda^2-2(1+\rho)\lambda+3\rho=0$$

So,
$$\lambda = \frac{2(1+\rho) \pm \sqrt{4(1+\rho)^2 - 12\rho}}{2} = (1+\rho) \left[1 \pm \sqrt{1 - \frac{3\rho}{(1+\rho)^2}} \right]$$

(iv) For $\rho=2$, $\lambda = 3 \left[1 \pm \sqrt{1 - \frac{6}{9}} \right] = 3 \left[1 \pm \sqrt{\frac{1}{3}} \right] = 3 \pm \sqrt{3}$

(v) Solving for the eigenvectors using $\begin{bmatrix} 2-\lambda & -1 \\ -\rho & 2\rho-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$

gives, for $\lambda = 3 - \sqrt{3}$, using the “top line” of the matrix equation

$$(2-3+\sqrt{3})x_1 - x_2 = 0 \text{ and so } x_2 = (\sqrt{3}-1)x_1$$

(As usual, we would get the same answer if we used the “bottom line” of the matrix equation.) We can write this as a vector $\begin{bmatrix} 1 \\ \sqrt{3}-1 \end{bmatrix}$.

For $\lambda = 3 + \sqrt{3}$, again using the “top line” of the matrix equation

$$(2-3-\sqrt{3})x_1 - x_2 = 0 \text{ and so } x_2 = -(1+\sqrt{3})x_1$$

which we can write as a vector $\begin{bmatrix} 1 \\ -(1+\sqrt{3}) \end{bmatrix}$.

(vi) Taking the inner product of these two vectors gives

$$\begin{bmatrix} 1 & -(1+\sqrt{3}) \end{bmatrix} \begin{bmatrix} 1 \\ \sqrt{3}-1 \end{bmatrix} = 1 + (1+\sqrt{3})(1-\sqrt{3}) = 1+1-3 = 1 \neq 0$$

So, these eigenvectors are *not* orthogonal. (In fact, eigenvectors of non-Hermitian operators are indeed not guaranteed to be orthogonal.)

Chapter 3

3.2.2

The energy of a photon is given by $E = \hbar\omega = hf$. Electromagnetic radiation travels at the speed of light, meaning that $f = c / \lambda$ for wavelength λ . This gives us $E = hc / \lambda$. So, $\lambda = hc / E$. One Rydberg is equivalent to $13.6 \text{ eV} \approx 2.17896 \times 10^{-18} \text{ J}$. Also, $c \approx 2.998 \times 10^8 \text{ m/s}$ and $h \approx 6.626 \times 10^{-34} \text{ J}\cdot\text{s}$. So, we obtain $\lambda = 91.2 \text{ nm}$, which is approximately at the border between x-rays and ultraviolet radiation (definitions of these spectral regions vary).

3.2.3

(i) Converting from potential to kinetic energy, we can calculate the momentum of the electron, which is then used to find its corresponding wavelength. So, with the magnitude of the electronic charge as e , and a voltage V (here 3V), we have $qV = (1/2)mv^2 \equiv p^2 / 2m$, so $p = \sqrt{2mqV}$. Hence

$$\lambda = h / p = \frac{h}{\sqrt{2mqV}} \approx \frac{6.626 \times 10^{-34} \text{ J}\cdot\text{s}}{\sqrt{2 \times 9.11 \times 10^{-31} \times 1.602 \times 10^{-19} \times 3}} \\ \approx 7.08 \times 10^{-10} \text{ m} = 0.708 \text{ nm} = 7.08 \text{ \AA}$$

Note: In this problem, the velocity of the electron is $\sim 1 \times 10^6 \text{ m/s}$, which is much less than the velocity of light $c \approx 3 \times 10^8 \text{ m/s}$, so it is still valid to use ordinary non-relativistic classical mechanics here.

(ii) As in the previous problem, we know that the wavelength of a photon is related to energy as $\lambda = hc / E \approx (1240 \text{ eV}\cdot\text{nm}) / E$. Thus, with an energy of 3 eV, the wavelength of the photon is approximately 413.3 nm.

3.3.1

We use Schrödinger's equation in one dimension, $\left(-\frac{\hbar^2}{2m} \frac{d^2}{dz^2} + V(z) \right) \psi = E\psi$.

From the question, $\psi(z) = Az \exp(-z^2 / L^2)$, with energy $E = 0$. We substitute this wavefunction into Schrödinger's equation. After performing the first of the two spatial derivatives, we have

$$-\frac{\hbar^2}{2m} \frac{d}{dz} \left[A \exp\left(\frac{-z^2}{L^2}\right) + Az \left(\frac{-2z}{L^2}\right) \exp\left(\frac{-z^2}{L^2}\right) \right] + V(x) Az \left(\frac{-z^2}{L^2}\right) = 0$$

Now performing the second of the spatial derivatives, we have

$$-\frac{\hbar^2}{2m} \left[\left(\frac{-2Az}{L^2}\right) \exp\left(\frac{-z^2}{L^2}\right) + \left(\frac{-4Az}{L^2}\right) \exp\left(\frac{-z^2}{L^2}\right) + \frac{4Az^3}{L^4} \exp\left(\frac{-z^2}{L^2}\right) \right] + V(x) Az \left(\frac{-z^2}{L^2}\right) = 0$$

Grouping terms using the definition given for the wavefunction, we have

$$-\frac{\hbar^2}{2m} \left(\frac{-2}{L^2} + \frac{-4}{L^2} + \frac{4z^2}{L^4} \right) \psi(z) + V(x) \psi(z) = 0$$

$$\text{Thus, } V(x) = \frac{2\hbar^2}{mL^4} \left(-\frac{3L^2}{2} + z^2 \right).$$

Note, incidentally, that since the origin for potential energy is always arbitrary, there is no problem with this potential energy being negative in some regions of space and no problem with the overall energy of this eigenstate being zero.

3.5.1

(i) For a particle in an infinitely deep potential well (1D), the energy levels have the form $E_n = \frac{\hbar^2}{2m} \left(\frac{n\pi}{L_z} \right)^2$ (where $n = 1, 2, 3, \dots$).

The zero-point energy of an electron in this well is equal to the energy of the first level

$$\begin{aligned} E_1 &= \frac{\hbar^2 \pi^2}{2mL_z^2} \simeq \frac{(1.055 \times 10^{-34} \text{ J}\cdot\text{s})^2 \pi^2}{2(9.11 \times 10^{-31} \text{ kg})(1.5 \times 10^{-9} \text{ m})^2} \\ &\simeq (2.6796 \times 10^{-20} \text{ J}) / (1.602 \times 10^{-19} \text{ J/eV}) \simeq 0.167 \text{ eV} \end{aligned}$$

(ii) The energy separation between the first and second energy eigenstates is

$$E_2 - E_1 = \frac{\hbar^2 \pi^2}{2mL_z^2} (2^2 - 1^2) = 3E_1 \simeq 3 \times (0.167 \text{ eV}) \simeq 0.501 \text{ eV}$$

3.5.6

$$\text{a) } \int_{-L}^L |\psi(z)|^2 dz = A^2 \int_{-L}^L \exp(ikz) \exp(-ikz) dz = A^2 z \Big|_{-L}^L = A^2 2L = 1. \text{ So, } A = \sqrt{\frac{1}{2L}}$$

$$\begin{aligned} \text{b) } \int_{-\infty}^{\infty} |\psi(z)|^2 dz &= A^2 \int_{-\infty}^{\infty} \exp(-2\kappa|z|) dz = 2A^2 \int_0^{\infty} \exp(-2\kappa z) dz \\ &= 2A^2 \frac{-1}{2\kappa} \exp(-2\kappa z) \Big|_0^{\infty} = 2A^2 \frac{-1}{2\kappa} (0 - 1) = \frac{A^2}{\kappa} = 1. \text{ So, } A = \sqrt{\kappa} \end{aligned}$$

$$\begin{aligned} \text{c) } \int_{-L/4}^{L/4} |\psi(z)|^2 dz &= A^2 \int_{-L/4}^{L/4} \cos^2 \left(\frac{2\pi z}{L} \right) dz = A^2 \int_{-L/4}^{L/4} \frac{1 + \cos(4\pi z / L)}{2} dz \\ &= A^2 \left(\frac{z}{2} - \frac{2\pi \sin(4\pi z / L)}{L} \right) \Big|_{-L/4}^{L/4} = A^2 \left(\frac{L}{4} \right) = 1. \text{ So, } A = \sqrt{\frac{4}{L}} \end{aligned}$$

3.5.7

a) $P(z) = |\psi(z)|^2 = (L/2) \sin^2(3\pi z / L)$. See Fig. 3.8. For $L/3 < z < 2L/3$, because this region is just the middle “bump” of three equal “bumps” in the probability density, we can write the answer $P_s = 1/3$ without necessarily having to perform the integral explicitly (though that will give the same answer).

$$\begin{aligned}
 \text{b) } P_s &= \left(\sqrt{\frac{2}{L}} \right)^2 \int_{0.8L}^L \sin^2 \left(\frac{2\pi z}{L} \right) dz = \frac{2}{L} \int_{0.8L}^L \frac{1}{2} \left[1 - \cos \left(\frac{4\pi z}{L} \right) \right] dz \\
 &= \frac{1}{L} \int_{0.8L}^L 1 dz - \frac{1}{L} \int_{0.8L}^L \frac{1}{2} \cos \left(\frac{4\pi z}{L} \right) dz = \frac{1}{L} [z]_{0.8L}^L - \frac{1}{L} \frac{L}{4\pi} \left[\sin \left(\frac{4\pi z}{L} \right) \right]_{0.8L}^L \\
 &= 0.2 - \frac{1}{4\pi} [0 - \sin(3.2\pi)] = 0.2 - \frac{0.588}{4\pi} = 0.153
 \end{aligned}$$

3.6.1

The Heisenberg uncertainty principle states in general that for any wavefunction ψ (which can be expressed as a distribution in either position or momentum space), the product of the variance in momentum and position has a minimum value given by $(\Delta p)(\Delta x) \geq \hbar / 2$, where “ Δ ” denotes the typical definition of statistical variance, i.e.,

$(\Delta p) = \sqrt{\langle p^2 \rangle - \langle p \rangle^2}$. For the purposes of this question, we substitute mass $\times$ velocity for momentum, giving $(m\Delta v)(\Delta x) \geq \hbar / 2$. The resulting solutions are:

$$(i) (\Delta v) \geq \frac{\hbar}{2m_e(\Delta x)} = \frac{1.055 \times 10^{-34} \text{ (J} \cdot \text{s)}}{2 \times 9.109 \times 10^{-31} \text{ (kg)} \times 10^{-9} \text{ (m)}} \approx 5790 \text{ m/s}$$

$$(ii) (\Delta v) \geq \frac{\hbar}{2m_p(\Delta x)} = \frac{1.055 \times 10^{-34} \text{ (J} \cdot \text{s)}}{2 \times 1.673 \times 10^{-27} \text{ (kg)} \times 10^{-9} \text{ (m)}} \approx 3.15 \text{ m/s}$$

$$(iii) (\Delta v) \geq \frac{\hbar}{2m_{cube}(\Delta x)} = \frac{1.055 \times 10^{-34} \text{ (Js)}}{2 \times \left((10^{-6})^3 \text{ (m}^3) \times 10,000 \left(\frac{\text{kg}}{\text{m}^3} \right) \right) \times 10^{-9} \text{ (m)}} \approx 5.28 \times 10^{-13} \text{ m/s}$$

3.6.3

We may follow the diagram and derivation in the text for a two-slit interference pattern. As shown in the text, $d_s = \lambda z_o / s$. It remains to solve for the wavelength given the energy in each case.

$$(i) \lambda = \frac{hc}{E} = \frac{4.136 \times 10^{-15} \text{ (eV} \cdot \text{s)} \times 3 \times 10^8 \left(\frac{\text{m}}{\text{s}} \right)}{1 \text{ (eV)}} = 1.24 \mu\text{m}$$

$$\text{So, } d = \frac{1.24(\mu\text{m}) \times 10^6(\mu\text{m})}{10(\mu\text{m})} = 124,000 \mu\text{m} = 0.124 \text{ m}$$

$$(ii) \lambda = \frac{h}{p} = \frac{h}{\sqrt{2mE}} = \frac{6.626 \times 10^{-34} \text{ (J} \cdot \text{s)}}{\sqrt{2 \times 9.109 \times 10^{-31} \text{ (kg)} \times 1.602 \times 10^{-19} \text{ (J)}}} = 1.23 \text{ nm}$$

$$\text{So, } d = \frac{0.00123(\mu\text{m}) \times 10^6(\mu\text{m})}{10(\mu\text{m})} = 123 \mu\text{m} = 0.000123 \text{ m}$$

3.7.3

Schrödinger's equation in the region $z > 0$ is $-\frac{\hbar^2}{2m_o} \frac{d^2\psi}{dz^2} + V_o\psi = E\psi$

or equivalently $-\frac{\hbar^2}{2m_o} \frac{d^2\psi}{dz^2} = (E - V_o)\psi$ where for this problem $(E - V_o) > 0$. The general solution to this equation is

$$\psi(z) = F \exp(ikz) + G \exp(-ikz) \text{ with } k = \sqrt{2m(E - V_o)/\hbar^2}$$

as can be verified by substitution. We are given that, in the region $z > 0$, there is no wave propagating to the “left” because there is no wave originating from the “right” side. This means that $G = 0$. Therefore,

$$\psi = F \exp(ikz)$$

with the above expression for k .

This is therefore a forward-propagating wave, though it will have a different (smaller) k than the wave on the left of the barrier because it corresponds to a smaller “kinetic energy”.

3.7.5

(i) For $z < 0$, we have a Schrödinger equation $-\frac{\hbar^2}{2m} \frac{\partial^2\psi(z)}{\partial z^2} = E\psi(z)$, which we can rewrite as $\frac{\partial^2\psi(z)}{\partial z^2} + \frac{2mE}{\hbar^2}\psi(z) = 0$, which has the general solution

$$\psi(z) = A \exp\left(i\sqrt{\frac{2mE}{\hbar^2}}z\right) + B \exp\left(-i\sqrt{\frac{2mE}{\hbar^2}}z\right), \text{ so } k = \sqrt{\frac{2mE}{\hbar^2}}.$$

For $0 < z < L_b$, we have a Schrödinger equation $-\frac{\hbar^2}{2m} \frac{\partial^2\psi(z)}{\partial z^2} + V_o\psi(z) = E\psi(z)$,

which we can rewrite as $\frac{\partial^2\psi(z)}{\partial z^2} - \frac{2m(V_o - E)}{\hbar^2}\psi(z) = 0$, which has the general

$$\text{solution } \psi(z) = C \exp\left(-\sqrt{\frac{2m(V_o - E)}{\hbar^2}}z\right) + D \exp\left(\sqrt{\frac{2m(V_o - E)}{\hbar^2}}z\right),$$

$$\text{so } \kappa = \sqrt{\frac{2m(V_o - E)}{\hbar^2}}.$$

(ii) Applying boundary conditions at $z = L_b$, i.e., $\psi_b(z = L_b) = \psi_{right}(z = L_b)$ and

$$\left.\frac{d\psi_b}{dz}\right|_{z=L_b} = \left.\frac{d\psi_{right}}{dz}\right|_{z=L_b}, \text{ gives, respectively,}$$

$$C \exp(-\kappa L_b) + D \exp(\kappa L_b) = F \text{ and } -\kappa C \exp(-\kappa L_b) + \kappa D \exp(\kappa L_b) = ikF \text{ and}$$

$$C = \frac{\kappa - ik}{2\kappa} \exp(\kappa L_b) F \text{ and } D = \frac{\kappa + ik}{2\kappa} \exp(-\kappa L_b) F \quad \text{Eqs. (1)}$$

Applying boundary conditions at $z = 0$, i.e., $\psi_{left}(z = 0) = \psi_b(z = 0)$ and

$$\left. \frac{d\psi_{left}}{dz} \right|_{z=0} = \left. \frac{d\psi_b}{dz} \right|_{z=0}, \text{ gives, respectively,}$$

$$A + B = C + D \text{ and } ikA - ikB = -\kappa C + \kappa D \quad \text{Eqs. (2)}$$

Eliminating B from Eqs. (2) gives $A = \frac{k + i\kappa}{2k} C + \frac{k - i\kappa}{2k} D$, and substituting from Eqs.

(1) for C and D gives

$$A = \frac{2k\kappa - i(k^2 - \kappa^2)}{4k\kappa} \exp(\kappa L_b) F + \frac{2k\kappa + i(k^2 - \kappa^2)}{4k\kappa} \exp(-\kappa L_b) F \quad \text{Eq. (3)}$$

From Eq. (3), we know A in terms of F and constants of the problem, so we can now write

$$\begin{aligned} T = |F|^2 / |A|^2 &= 1 / \left| \frac{\exp(\kappa L_b) + \exp(-\kappa L_b)}{2} - i \frac{k^2 - \kappa^2}{2k\kappa} \left[\frac{\exp(\kappa L_b) - \exp(-\kappa L_b)}{2} \right] \right|^2 \\ &= \frac{1}{\left| \cosh(\kappa L_b) - i \frac{k^2 - \kappa^2}{2k\kappa} \sinh(\kappa L_b) \right|^2} = \frac{1}{\cosh^2(\kappa L_b) + \left(\frac{k^2 - \kappa^2}{2k\kappa} \right)^2 \sinh^2(\kappa L_b)} \end{aligned}$$

(iii) Using $V_0 = 2\text{eV}$, $L_b = 0.25\text{nm}$, and $m \approx 9.1 \times 10^{-31}\text{kg}$ gives

$$k = 6.265\text{nm}^{-1}, \kappa = 3.617\text{nm}^{-1}$$

$$\text{So, } T = 1 / \left(\cosh^2(0.904) + \left(\frac{26.168}{45.321} \right)^2 \sinh^2(0.904) \right) = 0.413$$

Chapter 4

4.5.1

For the purposes of this problem, the positronium atom is just the hydrogen atom with a different value for the reduced mass μ , which we write as μ_p . Therefore, we can use expressions from the text.

$$\text{Reduced mass for positronium: } \frac{1}{\mu_p} = \frac{1}{m_e} + \frac{1}{m_e} \text{ so } \mu_p = \frac{m_e}{2}$$

$$\text{Bohr radius for positronium: } a_p = \frac{4\pi \epsilon_0 \hbar^2}{e^2 \mu} = \frac{8\pi \epsilon_0 \hbar^2}{e^2 m_e}$$

Rydberg constant for positronium: $Ry_p = \frac{\hbar^2}{2\mu a_0^2} = \frac{\hbar^2}{m_e} \left(\frac{e^2 m_e}{8\pi\epsilon_0 \hbar^2} \right)^2 = \frac{e^4 m_e}{64\pi^2 \epsilon_0^2 \hbar^2}$

(i) $E = \frac{-Ry_p}{n^2} = -\frac{e^4 m_e}{64\pi^2 \epsilon_0^2 \hbar^2 n^2}$

(ii) $|E_1| = \frac{e^4 m_e}{64\pi^2 \epsilon_0^2 \hbar^2} \approx \frac{(1.602 \times 10^{-19} \text{ C})^4 (9.109 \times 10^{-31} \text{ kg})}{64\pi^2 (8.854 \times 10^{-12} \text{ (F/m)})^2 (1.055 \times 10^{-34} \text{ J s})^2}$
 $\approx 1.089 \times 10^{-18} \text{ J} \equiv \frac{1 \text{ eV}}{1.602 \times 10^{-19} \text{ J}} \approx 6.8 \text{ eV}$

(iii) From above, $a_p = \frac{8\pi\epsilon_0 \hbar^2}{e^2 m_e}$

(iv) $a_p = \frac{8\pi(8.854 \times 10^{-12} \text{ (F/m)})(1.055 \times 10^{-34} \text{ J s})^2}{(1.602 \times 10^{-19} \text{ C})^2 (9.109 \times 10^{-31} \text{ kg})} = 1.06 \times 10^{-10} \text{ m} = 0.106 \text{ nm}$

So we see that the positronium atom is approximately twice as large as the hydrogen atom and has approximately half of the “binding energy” for the $n = 1$ state.

4.5.4

The eigenenergy of an electron orbital for the hydrogen atom is $E_{Hn} = -Ry / n^2$ where $Ry \approx 13.6 \text{ eV}$. (Note that, for the hydrogen atom, the energies are only dependent on n .) Therefore, the energy separation between two states is given by

$$\Delta E_{Hn} = -Ry \left(\frac{1}{n_2^2} - \frac{1}{n_1^2} \right)$$

(i) $n_2 = 3$ and $n_1 = 2$. Therefore, $\Delta E = 1.89 \text{ eV}$.

(ii) $n_2 = 6$ and $n_1 = 4$. Therefore, $\Delta E = 0.47 \text{ eV}$.

(iii) $n_2 = 2$ and $n_1 = 2$. Therefore, $\Delta E = 0 \text{ eV}$.

4.8.1

A fully occupied s orbital has 2 electrons and a fully occupied p orbital has 6 electrons. Atomic number = number of electrons = $2 + 2 + 6 + 2 = 12$. This atomic number corresponds to the element magnesium (Mg).

4.8.2

$n = 1, l = 0, 2$ electrons

$n = 2, l = 0, 2$ electrons

$n = 2, l = 1, 6$ electrons

$n = 3, l = 0, 2$ electrons

$n = 3, l = 1, 6$ electrons

$n = 4, l = 0, 1$ electron

4.13.3

Starting with the given wavefunction expression and simplifying, we obtain

$$\begin{aligned}\psi(z) &= \exp(ikz) \left[\exp\left(-i\frac{2\pi}{a}z\right) - \exp\left(i\frac{2\pi}{a}z\right) \right] = \exp(ikz) \left(-2i \sin\left(\frac{2\pi}{a}z\right) \right) \\ &= -2i \sin\left(\frac{2\pi}{a}z\right) \exp(ikz)\end{aligned}$$

This is in Bloch form with the choice $u(z) = -2i \sin\left(\frac{2\pi}{a}z\right)$ for the unit cell function.

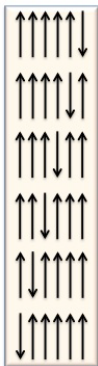
We can directly check that this function is the same in every unit cell. We consider the function in the “next” unit cell and show that by moving over by unit repeat length a , the value of the unit cell function is exactly the same, i.e.,

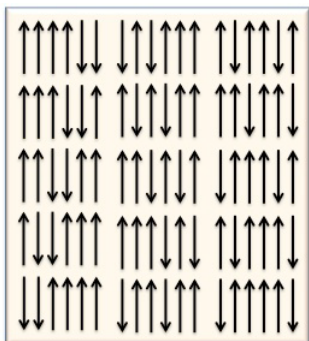
$$u(z+a) = -2i \sin\left(\frac{2\pi}{a}(z+a)\right) = -2i \sin\left(\frac{2\pi}{a}z + 2\pi\right) = -2i \sin\left(\frac{2\pi}{a}z\right) = u(z)$$

Chapter 5

5.2.1

Follow the same approach as in the text, the following table can be obtained.

Microstate	Macrostate	Multiplicity	Excess, $2s_e$
	6 up (“heads”)	$g = 1$	6
	5 up (“heads”) 1 down (“tails”)	$g = \frac{6!}{5!1!} = 6$	4

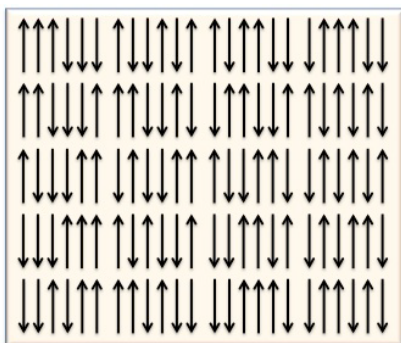


4 up ("heads")

2 down
("tails")

$$g = \frac{6!}{4!2!} = 15$$

2

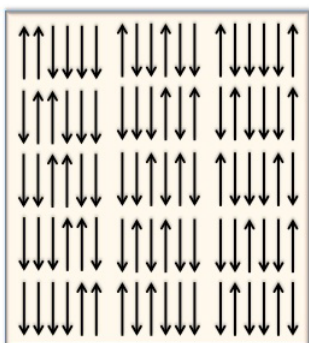


3 up ("heads")

3 down
("tails")

$$g = \frac{6!}{3!3!} = 20$$

0

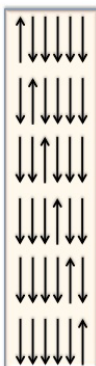


2 up ("heads")

4 down
("tails")

$$g = \frac{6!}{4!2!} = 15$$

-2



1 up ("heads")

5 down
("tails")

$$g = \frac{6!}{5!1!} = 6$$

-4

6 down
("tails")

$$g = 1$$

-6

From the table, we can tell the most probable macrostate is with 3 heads and 3 tails, i.e., $2s_e = 0$.

5.2.3

The results below are calculated for a coin that is tossed $N = 100$ times.

(i) For the case where only two of the results are “heads” and the others are all “tails”, the binomial distribution gives

$$g = \frac{N!}{(N-k)!k!} = \frac{100!}{(100-2)!2!} = \frac{100!}{98!2!} = \frac{(100)(99)}{2} = 4950$$

So, there are 4950 possible such sequences.

With equal numbers of heads and tails, necessarily the “excess” and s_e are both zero. According to the Gaussian approximation of the binomial distribution,

$$\begin{aligned} g(N, s_e) &\approx \sqrt{\frac{2}{\pi N}} 2^N \exp\left(-\frac{2s_e^2}{N}\right) = \sqrt{\frac{2}{100\pi}} 2^{100} \exp\left(-\frac{2 \times (0)^2}{100}\right) = \sqrt{\frac{1}{50\pi}} 2^{100} \times 1 \\ &\approx 1.01 \times 10^{29} \end{aligned}$$

So, there are approximately 1.01×10^{29} such possible sequences.

Note that this is massively larger than the number of sequences with only two “heads”.

5.3.1

a) The given energy corresponds to $s_e = -2$. The multiplicity of this state is 6 (as also in the solution of problem 5.2.1), while the whole system has $2^6 = 64$ different microstates. Therefore, the probability of the system having the given energy is $6/64 = 9.4\%$.

b) The expected energy term can be written as

$$\langle U \rangle = \sum_n P(s_e^n) U(s_e^n)$$

which can be expanded into

$$\langle U \rangle = -2\mu_s B \sum_{s_e=-3}^3 \binom{6}{3+s_e} s_e = -2\mu_s B (-3-12-15+0+15+12+3) = 0$$

5.4.3

Our answers are based on the defining relation $\frac{\Delta S}{\Delta U} = \frac{1}{T}$.

$$(a) \Delta S = \frac{\Delta U}{T} = \frac{3.34 \times 10^5 \text{ J}}{273 \text{ K}} = 1.22 \times 10^3 \frac{\text{J}}{\text{K}}$$

$$(b) \Delta S = \frac{\Delta U}{T} = \frac{-2.26 \times 10^6 \text{ J}}{373 \text{ K}} = -6.05 \times 10^3 \frac{\text{J}}{\text{K}}$$

5.5.3

We presume we extract an amount of heat of magnitude U_C from the cold reservoir at temperature T_C . So, the change in entropy of the cold reservoir is $\Delta S_C = -U_C / T_C$. The heat energy added to the high temperature reservoir, by conservation of energy, has to be of magnitude $U_H = U_C + W$, where W is the (positive) amount of work energy put into the heat pump engine. So, the change in entropy of the hot reservoir is $\Delta S_H = U_H / T_H$.

In this problem, we want to end up with a heat pump efficiency (η_{heatpump}) limit based on a definition of the heat added to the hot reservoir per unit work, so we want to cast the problem in terms of U_H and W . So, we rewrite ΔS_C as

$$\Delta S_C = -\frac{U_C}{T_C} = -\left(\frac{U_H - W}{T_C}\right) = \frac{W - U_H}{T_C}$$

The total entropy cannot decrease, so $\Delta S_C + \Delta S_H \geq 0$.

So,

$$\frac{W - U_H}{T_C} + \frac{U_H}{T_H} \geq 0$$

So,

$$U_H \left(\frac{1}{T_H} - \frac{1}{T_C} \right) \geq -\frac{W}{T_C}, \text{ i.e., } U_H \left(\frac{1}{T_C} - \frac{1}{T_H} \right) \leq \frac{W}{T_C}$$

So,

$$\frac{U_H}{W} \left(1 - \frac{T_C}{T_H} \right) \leq 1$$

i.e.,

$$\eta_{\text{heatpump}} \equiv \frac{U_H}{W} \leq \frac{T_H}{T_H - T_C}$$

(ii) With these numbers

$$\eta_{\text{heatpump}} \leq \frac{T_H}{T_H - T_C} = \frac{298.15}{298.15 - 273.15} = \frac{298.15}{25} \approx 11.9 \equiv 1109\%$$

Indeed, a heat pump can be theoretically much more than 100% efficient, and it is one of the possible technological approaches to heating a house efficiently. As the outside temperature gets colder, however, it does get less efficient.

5.6.1

The energy difference between the two levels is $\varepsilon = -Ry(1/36 - 1/25) = 0.166 \text{ eV}$. Using the Boltzmann factor $\exp\left(-\frac{\varepsilon}{k_B T}\right)$ we obtain the relative occupation probabilities:

a) 1.4×10^{-3}

b) 0.716

At higher temperatures, the likelihood of populating a higher energy level over a lower one increases. While at room temperatures the occupation of a higher level is

diminishingly small, at temperatures as high as the Sun temperature the occupations of these two levels are quite similar.

5.8.2

$k_B T$ is approximately 25.85 meV at 300 K. Electrons follow the Fermi-Dirac distribution. For an occupation probability of 10^{-2} , $0.01 = 1 / \left[\exp\left(\frac{E_C - E_F}{25.85 \text{ meV}}\right) + 1 \right]$.

So, $0.01 \exp\left(\frac{E_C - E_F}{25.85 \text{ meV}}\right) + 0.01 = 1$. Hence, $\frac{E_C - E_F}{25.85 \text{ meV}} = \ln\left(\frac{0.99}{0.01}\right)$, which gives $E_C - E_F = 119 \text{ meV}$.

So, the Fermi level lies in the band gap, 119 meV below the bottom of the conduction band. Note that the Fermi level does not have to correspond to an allowed energy state of the system.

Chapter 6

6.4.4

The magnitude of the current flowing through the diode is given by the difference between the diffusion current and the reverse “leakage” current $I_S = 1 \text{ nA}$. This is given

by the equation
$$I = I_S \left[\exp\left(\frac{eV}{k_B T}\right) - 1 \right]$$

where here we take $k_B T = 25 \text{ meV}$. Substituting in each value of voltage V will give the current flowing through the diode. Negative values correspond to the “backward” current direction and positive values correspond to the “forward” direction.

(i) $I = I_S \left(\exp\left(\frac{e(-25 \text{ mV})}{k_B T}\right) - 1 \right) = -0.62 \text{ nA}$. So $|I| = 0.62 \text{ nA}$ in the reverse direction

(ii) $I = I_S \left(\exp\left(\frac{e(0 \text{ mV})}{k_B T}\right) - 1 \right) = 0 \text{ nA}$. So $|I| = 0 \text{ nA}$, which is zero current (with, therefore, no direction)

(iii) $I = I_S \left(\exp\left(\frac{e(25 \text{ mV})}{k_B T}\right) - 1 \right) = 1.63 \text{ nA}$. So $|I| = 1.63 \text{ nA}$ in the forward direction

(iv) $I = I_S \left(\exp\left(\frac{e(100 \text{ mV})}{k_B T}\right) - 1 \right) = 46.8 \text{ nA}$. So $|I| = 46.8 \text{ nA}$ in the forward direction

(v) $I = I_S \left(\exp\left(\frac{e(500 \text{ mV})}{k_B T}\right) - 1 \right) = 250 \text{ nA}$. So $|I| = 250 \text{ nA}$ in the forward direction

Chapter 7

7.4.1

The box can be treated as a perfect black body since light will be trapped inside the box due to the perfectly reflecting walls. According to the Stefan-Boltzmann law, the energy

per unit volume inside the black body is $U = \frac{\pi^2 k_B^4}{15 \hbar^3 c^3} T^4$

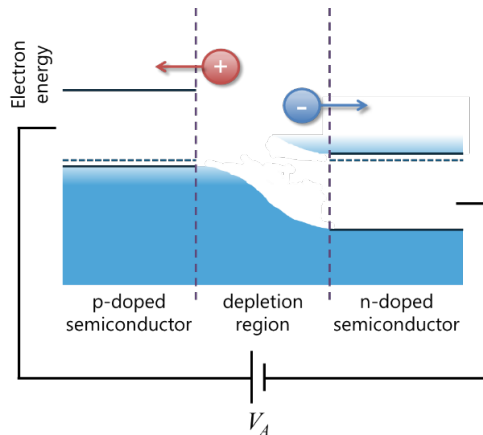
So, the total energy stored inside the box is

$$E = \frac{3.14^2 \times (1.38 \times 10^{-23} \text{ J/K})^4}{15 \times (1.05 \times 10^{-34} \text{ J} \cdot \text{s})^3 \times (3 \times 10^8 \text{ m/s})^3} \times (5800 \text{ K})^4 \times 1 \text{ m}^3 = 0.863 \text{ J}$$

Chapter 8

8.2.1

a) Holes move to the left and electrons move to the right. Both are moving in the opposite directions compared to how they should move for a forward current. (For example, a forward current of holes would be moving to the right in the diagram below.) Thus, the photogenerated electrical current is negative. The driving force of the movement is the “built-in” electric field inside the depletion region.



b) From part (a), we know the photogenerated carriers give negative (“reverse”) current. To reach zero current, we must apply a bias that generates positive current to cancel out the negative current. Therefore, we must forward bias the diode, i.e. $V_A > 0$.

Appendix F

Online videos

F.1 Modern Physics for Engineers and Scientists videos and slides

Video lectures are available for all the material in Chapters 1 through 9 of this book. These videos are listed below in sequence. Two lists are given. First, a short list gives just the lecture titles and textbook section references (“MPES” refers to this text, *Modern Physics for Engineers and Scientists*). In an electronic version of this text, the title text is a live link to the corresponding lecture video. Second, a full list includes short abstracts and keywords for each lecture part, together with explicit links to both the videos and the associated lecture slides, and QR codes for ease of linking when using a paper version of this book. The longer list may also help in finding or searching for specific topics and referencing or linking parts of the course content.

F.2 Short list of lectures and slides

Lecture 1 Introduction and background

Lecture 1a - [Introduction to the class](#) MPES 1.1

Lecture 1b - [The background to modern physics](#) MPES 1.2

Lecture 1c - [Transitioning to modern physics](#) MPES 1.3

Lecture 2 Oscillations and waves 1

Lecture 2a - [Classical mechanics](#) MPES 2.2

Lecture 2b - [Modes](#) MPES 2.3

Lecture 2c - [Simple harmonic oscillator](#) MPES 2.4

Lecture 3 Oscillations and waves 2

Lecture 3a - [Eigen equations and operators](#) MPES 2.4

Lecture 3b - [The classical wave equation](#) MPES 2.5

Lecture 3c - [The Helmholtz equation](#) MPES 2.6

Lecture 3d - [Standing waves](#) MPES 2.7

Lecture 4 Oscillations and waves 3

Lecture 4a - [A coupled oscillator](#) MPES 2.8

Lecture 4b - [Inner products, orthogonality, and basis sets](#) MPES 2.9

Lecture 4c - [Hermitian operators and sets of functions](#) MPES 2.9

Lecture 5 The quantum view of the world 1

Lecture 5a - [The beginning of quantum mechanics](#) MPES 3.1

Lecture 5b - [Electrons and atoms](#) MPES 3.2

Lecture 6 The quantum view of the world 2

Lecture 6a - [Electrons and waves](#) MPES 3.2, 3.3

Lecture 6b - [Solving Schrödinger's equation – a particle in a box](#) MPES 3.5

Lecture 6c - [Normalization and probability](#) MPES 3.5

Lecture 6d - [The nature of the particle-in-a-box solutions](#) MPES 3.5

Lecture 7 The quantum view of the world 3

Lecture 7a - [Waves, diffraction and uncertainty](#) MPES 3.6

Lecture 7b - [Diffraction by two slits – Young's slits](#) MPES 3.6

Lecture 7c - [Young's slits and quantum mechanics](#) MPES 3.6

Lecture 8 The quantum view of the world 4

Lecture 8a - [The nature of quantum mechanical particles](#) MPES 3.6

Lecture 8b - [Waves and measurement](#) MPES 3.6

Lecture 9 The quantum view of the world 5

Lecture 9a - [Tunneling](#) MPES 3.7

Lecture 9b - [Solving for barriers of finite height](#) MPES 3.7

Lecture 9c - [Tunneling through a barrier](#) MPES 3.7

Lecture 10 Particles, atoms, and crystals 1

- Lecture 10a - [The hydrogen atom and center-of-mass coordinates](#) MPES 4.1-4.3
Lecture 10b - [The hydrogen atom solutions and angular behavior](#) MPES 4.3-4.4
Lecture 10c - [Spherical harmonics for a classical problem](#) MPES 4.4
Lecture 10d - [Polar plots of spherical harmonics](#) MPES 4.4
Lecture 10e - [Spherical harmonics and atomic orbitals](#) MPES 4.4
-

Lecture 11 Particles, atoms, and crystals 2

- Lecture 11a - [Hydrogen atom radial solutions](#) MPES 4.5
Lecture 11b - [Hydrogen atom complete solutions](#) MPES 4.5-4.6
Lecture 11c - [Electron spin and Pauli exclusion](#) MPES 4.7
-

Lecture 12 Particles, atoms, and crystals 3

- Lecture 12a - [Many-electron atoms](#) MPES 4.8
Lecture 12b - [Filling “shells” in atoms](#) MPES 4.8
Lecture 12c - [Fermions and bosons](#) MPES 4.9
Lecture 12d - [Identical particles](#) MPES 4.9
-

Lecture 13 Particles, atoms, and crystals 4

- Lecture 13a - [Coupled systems](#) MPES 4.10
Lecture 13b - [Crystals](#) MPES 4.11
Lecture 13c - [Emergence of bands](#) MPES 4.12
-

Lecture 14 Particles, atoms, and crystals 5

- Lecture 14a - [Band structures in crystals](#) MPES 4.13
Lecture 14b - [The Bloch theorem](#) MPES 4.13
Lecture 14c - [Solving with periodic boundary conditions](#) MPES 4.13
-

Lecture 15 Particles, atoms, and crystals 6

- Lecture 15a - [Band structures](#) MPES 4.14
Lecture 15b - [Crystal momentum and effective mass](#) MPES 4.14
Lecture 15c - [Band structures in three dimensions](#) MPES 4.15
-

Lecture 15d - [Plotting actual band structures](#) MPES 4.15

Lecture 15e - [Metals, semiconductors and insulators](#) MPES 4.16

Lecture 16 Thermal distributions 1

Lecture 16a - [Tossing coins, microstates, and macrostates](#) MPES 5.2

Lecture 16b - [Binomial distribution and Stirling's approximation](#) MPES 5.2

Lecture 16c - [Two-state spin systems and microstates](#) MPES 5.3

Lecture 17 Thermal distributions 2

Lecture 17a - [Systems in thermal contact](#) MPES 5.3

Lecture 17b - [Maximizing multiplicities for spin systems](#) MPES 5.3

Lecture 17c - [Maximizing multiplicity for general systems](#) MPES 5.4

Lecture 18 Thermal distributions 3

Lecture 18a - [Entropy and temperature](#) MPES 5.4

Lecture 18b - [Entropy and heat flow](#) MPES 5.4

Lecture 18c - [Carnot efficiency limit for heat engines](#) MPES 5.5

Lecture 19 Thermal distributions 4

Lecture 19a - [The Boltzmann factor](#) MPES 5.6

Lecture 19b - [Chemical potential](#) MPES 5.7

Lecture 19c - [Chemical potential and the Gibbs factor](#) MPES 5.7

Lecture 20 Thermal distributions 5

Lecture 20a - [The Fermi-Dirac distribution](#) MPES 5.8

Lecture 20b - [The Bose-Einstein and Planck distributions](#) MPES 5.8

Lecture 20c - [The Maxwell-Boltzmann distribution](#) MPES 5.8

Lecture 21 Bands and electronic devices

Lecture 21a - [Electrons and holes in bands](#) MPES 6.2

Lecture 21b - [Semiconductor doping and diodes](#) MPES 6.3, 6.4

Lecture 21c - [Voltages and Fermi levels](#) MPES 6.4 (and 5.7 for equivalent chemical potential definition)

Lecture 21d - [Biasing semiconductor devices](#) MPES 6.4

Lecture 22 Light and quantum mechanics 1

Lecture 22a - [Light, the photoelectric effect and the photon](#) MPES 7.2

Lecture 22b - [Light and modes](#) MPES 7.3

Lecture 22c - [Thermal radiation](#) MPES 7.4

Lecture 23 Light and quantum mechanics 2

Lecture 23a - [Black body radiation and Kirchhoff's law](#) MPES 7.5

Lecture 23b - [Einstein's A and B coefficient argument](#) MPES 7.6

Lecture 24 Semiconductor optoelectronics 1

Lecture 24a - [Optoelectronic devices and photodetectors](#) MPES 8.1, 8.2

Lecture 24b - [Light emission](#) MPES 8.3

Lecture 25 Semiconductor optoelectronics 2

Lecture 25a - [Absorption and emission processes](#) MPES 8.3

Lecture 25b - [Lasers](#) MPES 8.3

Lecture 25c - [Epilogue](#) MPES Chapter 9

F.3 Full list of lectures and slides

Lecture 1 Introduction and background

Lecture 1a - Introduction to the class MPES 1.1



This lecture introduces the motivation for why we need ideas of modern physics so we can understand the world around us and engineer the technologies that make it work. It summarizes the topics to be taught in later lectures, including quantum mechanics and statistical mechanics.

Specific topics include the quantum view of the world, particles, atoms and crystals, thermal distributions, bands and electronic devices, light and quantum mechanics, and semiconductor optoelectronics.

Keywords: Physics, Quantum physics, Statistical mechanics, Crystalline semiconductors, Quantum optics, Optoelectronic devices

Video: <https://purl.stanford.edu/rs056yh7155> DOI: <https://doi.org/10.25740/rs056yh7155>

Slides: <https://purl.stanford.edu/sq388rv0716> DOI: <https://doi.org/10.25740/sq388rv0716>

Lecture 1b - The background to modern physics MPES 1.2



This lecture part summarizes the background to modern physics. It starts with early ideas of matter, through the development of the scientific method and its consequences for understanding of matter, and continues with the development of laws of motion. It then introduces the development of physical concepts of light and electromagnetism through to Maxwell's equations. Finally, it summarizes the history of the understanding of heat and thermodynamics, up to the ideas of the second law of thermodynamics and entropy.

Keywords: History of science, Matter, Laws of motion, Heat, Thermodynamics, Entropy

Video: <https://purl.stanford.edu/pd853tx7523> DOI: <https://doi.org/10.25740/pd853tx7523>

Slides: <https://purl.stanford.edu/vb612ph6646> DOI: <https://doi.org/10.25740/vb612ph6646>

Lecture 1c - Transitioning to modern physics MPES 1.3



This lecture part summarizes the state of our knowledge of what we could call "classical" physics as of about 1870, and our ability to exploit it, including electromagnetism and thermodynamics, and the beginning of the transition to the ideas of modern physics, including the basis of chemistry, the physical processes of light, and what lies behind thermodynamics.

Keywords: Classical physics, Light, Thermodynamics, Chemistry

Video: <https://purl.stanford.edu/qm226hf2245> DOI: <https://doi.org/10.25740/qm226hf2245>

Slides: <https://purl.stanford.edu/jy074vx1575> DOI: <https://doi.org/10.25740/jy074vx1575>

Lecture 2 Oscillations and waves 1

Lecture 2a - Classical mechanics MPES 2.2



This lecture part reminds us of classical mechanics ideas like kinetic energy, momentum, Newton's second law, potential energy and force, and the relations between these.

Keywords: Kinetic energy, Momentum, Newton's second law, Potential energy, Force

Video: <https://purl.stanford.edu/bf612xk1137> DOI: <https://doi.org/10.25740/bf612xk1137>

Slides: <https://purl.stanford.edu/vr422dz4554> DOI: <https://doi.org/10.25740/vr422dz4554>

Lecture 2b - Modes MPES 2.3



This lecture part introduces the idea of modes, including simple physical oscillating modes. It emphasizes their generality across many areas of physics and mathematics, including acoustics, mechanical and

electrical engineering, quantum mechanics. It introduces examples from musical instruments and vibrations, including standing waves.

Keywords: Modes, Standing waves, Oscillations in musical instruments

Video: <https://purl.stanford.edu/vr438tv1876> DOI: <https://doi.org/10.25740/vr438tv1876>

Slides: <https://purl.stanford.edu/sy173zy8535> DOI: <https://doi.org/10.25740/sy173zy8535>

Lecture 2c - Simple harmonic oscillator MPES 2.4



This lecture part introduces the physics and mathematical description of a simple harmonic oscillator, as in a mass on a spring.

Keywords: Simple harmonic oscillator, Oscillator

Video: <https://purl.stanford.edu/cy938rs1418> DOI: <https://doi.org/10.25740/cy938rs1418>

Slides: <https://purl.stanford.edu/tg077df3128> DOI: <https://doi.org/10.25740/tg077df3128>

Lecture 3 Oscillations and waves 2

Lecture 3a - Eigen equations and operators MPES 2.4



Starting with the simple harmonic oscillator as an example, this lecture part introduces the ideas of linear operators and eigen equations.

Keywords: Linear operators, Eigen equations, Simple harmonic oscillator

Video: <https://purl.stanford.edu/dy414rr3745> DOI: <https://doi.org/10.25740/dy414rr3745>

Slides: <https://purl.stanford.edu/bt435nk2445> DOI: <https://doi.org/10.25740/bt435nk2445>

Lecture 3b - The classical wave equation MPES 2.5



This lecture part introduces and derives the classical wave equation, as for a wave on a string

Keywords: Wave equation, Wave on a string

Video: <https://purl.stanford.edu/sr153qs8657> DOI: <https://doi.org/10.25740/sr153qs8657>

Slides: <https://purl.stanford.edu/tv859dj3309> DOI: <https://doi.org/10.25740/tv859dj3309>

Lecture 3c - The Helmholtz equation MPES 2.6



This lecture part introduces the Helmholtz wave equation, which is the classical wave equation for one specific frequency.

Keywords: Helmholtz wave equation, Wave equation

Video: <https://purl.stanford.edu/hf639rj0381> DOI: <https://doi.org/10.25740/hf639rj0381>

Slides: <https://purl.stanford.edu/cp506cb7154> DOI: <https://doi.org/10.25740/cp506cb7154>

Lecture 3d - Standing waves MPES 2.7



This lecture part introduces standing waves on a string stretched between two supports. This also introduces the idea of boundary conditions when solving differential equations.

Keywords: Standing waves, Boundary conditions (Differential equations)

Video: <https://purl.stanford.edu/cj689wp2705> DOI: <https://doi.org/10.25740/cj689wp2705>

Slides: <https://purl.stanford.edu/kd448wh1664> DOI: <https://doi.org/10.25740/kd448wh1664>

Lecture 4 Oscillations and waves 3

Lecture 4a - A coupled oscillator MPES 2.8



This lecture part introduces coupled oscillating systems, using the example of two masses connected by springs in a line between two supports, deriving the resulting oscillating modes as good examples of eigen modes or eigen functions, and preparing for more complicated coupled systems in quantum mechanics and elsewhere.

Keywords: Coupled oscillator

Video: <https://purl.stanford.edu/gj820np3673> DOI: <https://doi.org/10.25740/gj820np3673>

Slides: <https://purl.stanford.edu/mn622wd9822> DOI: <https://doi.org/10.25740/mn622wd9822>

Lecture 4b - Inner products, orthogonality, and basis sets MPES 2.9



This lecture part introduces core mathematical ideas needed to understand oscillating modes and, later, quantum mechanical states. Specifically, it introduces linear algebra ideas of inner products, which generally also define the idea of orthogonality, and basis sets of functions that can be used to describe other functions.

Keywords: Inner product, Orthogonality, Basis sets

Video: <https://purl.stanford.edu/vq983jh7795> DOI: <https://doi.org/10.25740/vq983jh7795>

Slides: <https://purl.stanford.edu/rv182yy1324> DOI: <https://doi.org/10.25740/rv182yy1324>

Lecture 4c - Hermitian operators and sets of functions MPES 2.9



This lecture part starts by introducing key ideas for working with complex matrices, especially the ideas of Hermitian adjoints and Hermitian matrices or, more generally, Hermitian operators. Such Hermitian operators are very useful for describing many physical systems, such as simple oscillators, and also later quantum mechanical systems. Among other properties, their eigenfunctions are orthogonal, and can represent, for example, the modes of many oscillating systems. Such eigenfunctions

also generate mathematically complete and orthogonal basis sets for describing their corresponding physical systems - a very powerful property.

Keywords: Linear operators, Hermitian adjoint, Hermitian matrix, Complete sets

Video: <https://purl.stanford.edu/pb380yd0294> DOI: <https://doi.org/10.25740/pb380yd0294>

Slides: <https://purl.stanford.edu/sn311ft9797> DOI: <https://doi.org/10.25740/sn311ft9797>

Lecture 5 The quantum view of the world 1

Lecture 5a - The beginning of quantum mechanics MPES 3.1



This lecture part starts by summarizing many of the puzzles in physics towards the end of the 19th century, including in particular the form in color or wavelength of the light from the sun, electric light bulbs, or more generally the light emission from a hot object if it is perfectly "black" or absorbing at all wavelengths. The resolution of this problem by Max Planck became the start of modern quantum mechanics.

Keywords: Quantum mechanics, Spectrum of hot objects

Video: <https://purl.stanford.edu/wr173vx1775> DOI: <https://doi.org/10.25740/wr173vx1775>

Slides: <https://purl.stanford.edu/sx884mb7664> DOI: <https://doi.org/10.25740/sx884mb7664>

Lecture 5b - Electrons and atoms MPES 3.2



This lecture part summarizes the early ideas to understand the quantum mechanics of atoms, including the Bohr model that, though not correct, gave a first quantum mechanical explanation of aspects of atomic emission lines, and helped lead towards the later correct models.

Keywords: Bohr model

Video: <https://purl.stanford.edu/yy363hk6055> DOI: <https://doi.org/10.25740/yy363hk6055>

Slides: <https://purl.stanford.edu/cp824hf8994> DOI: <https://doi.org/10.25740/cp824hf8994>

Lecture 6 The quantum view of the world 2

Lecture 6a - Electrons and waves MPES 3.2, 3.3



This lecture part introduces the idea of electrons as waves, starting with de Broglie's hypothesis, and then leading into Schrodinger's (time-independent) wave equation, briefly summarizing also the subsequent early evidence for electrons behaving as waves.

Keywords: de Broglie hypothesis, Electron diffraction, Schrödinger equation

Video: <https://purl.stanford.edu/kc067pb5569> DOI: <https://doi.org/10.25740/kc067pb5569>

Slides: <https://purl.stanford.edu/mn915cg4923> DOI: <https://doi.org/10.25740/mn915cg4923>

Lecture 6b - Solving Schrödinger's equation – a particle in a box MPES 3.5



This lecture part introduces and solves the simple quantum mechanical problem of a particle in a box, illustrating how solving Schrödinger's equation leads to discrete energy levels and eigen functions that we can think of as wavefunctions.

Keywords: Particle in a box, Schrödinger equation

Video: <https://purl.stanford.edu/sz213qv3851> DOI: <https://doi.org/10.25740/sz213qv3851>

Slides: <https://purl.stanford.edu/vs499pp9532> DOI: <https://doi.org/10.25740/vs499pp9532>

Lecture 6c - Normalization and probability MPES 3.5



This lecture part introduces the idea of "normalization" of a wavefunction, which then means that the modulus squared of the wavefunction directly gives the probability of finding the particle in that vicinity (strictly, the probability density)

Keywords: Normalization of the wavefunction, Probability density

Video: <https://purl.stanford.edu/xt101qb9456> DOI: <https://doi.org/10.25740/xt101qb9456>

Slides: <https://purl.stanford.edu/jz515cg5099> DOI: <https://doi.org/10.25740/jz515cg5099>

Lecture 6d - The nature of the particle-in-a-box solutions MPES 3.5



This lecture part uses the solutions to the particle-in-a-box problem to illustrate several general behaviors found in quantum mechanics, including the notion of only specific allowed energy "states", quantum numbers that label these states, the notion of "parity" or odd or even symmetry of wavefunctions, and other quantum mechanical behaviors, including "zero-point energy", and points in space that the electron in a given state will never be found. This particle-in-a-box problem allows simple estimates of the energies of quantum mechanical electron states when the electron is "quantum-confined" in a small space.

Keywords: Particle in a box, Zero-point energy, Parity of a wavefunction, Quantum confinement

Video: <https://purl.stanford.edu/qr141gt0040> DOI: <https://doi.org/10.25740/qr141gt0040>

Slides: <https://purl.stanford.edu/fr754sk2761> DOI: <https://doi.org/10.25740/fr754sk2761>

Lecture 7 The quantum view of the world 3

Lecture 7a - Waves, diffraction and uncertainty MPES 3.6



This lecture part introduces the idea of uncertainty principles, first by using a classical uncertainty principle common for most kinds of waves - the (inverse) relation between the size of an aperture and the diffraction angle or "spread" of the wave. When applied to quantum mechanical waves, we can justify the Heisenberg uncertainty principle. The meaning of the uncertainty principle in quantum mechanics is often unfortunately mixed up with the difficult issue of measurement in quantum mechanics and the "collapse of the wavefunction".

Keywords: Heisenberg uncertainty principle, Diffraction angle, Collapse of the wavefunction, Quantum mechanical measurement

Video: <https://purl.stanford.edu/gt525kd3441> DOI: <https://doi.org/10.25740/gt525kd3441>

Slides: <https://purl.stanford.edu/cg993bh9578> DOI: <https://doi.org/10.25740/cg993bh9578>

Lecture 7b - Diffraction by two slits – Young's slits MPES 3.6



This lecture part introduces the classical wave behavior of Young's slits - two closely spaced slits in a mask that lead to a diffraction pattern behind the mask that, among other things, verifies light is a wave and allows a measurement of its wavelength.

Keywords: Young's slits, Diffraction

Video: <https://purl.stanford.edu/hj539ht2308> DOI: <https://doi.org/10.25740/hj539ht2308>

Slides: <https://purl.stanford.edu/wm946hq0961> DOI: <https://doi.org/10.25740/wm946hq0961>

Lecture 7c - Young's slits and quantum mechanics MPES 3.6



This lecture part discusses a classic conundrum in quantum mechanics that is illustrated by Young's slits. Electrons do diffract as expected by Young's slits to give an interference pattern, but we apparently cannot "know" which slit the electron went through. In quantum mechanics, in one view that question is meaningless; indeed if we do measure that, we lose the interference pattern.

Keywords: Quantum mechanical measurement, Young's slits

Video: <https://purl.stanford.edu/xg100gf3909> DOI: <https://doi.org/10.25740/xg100gf3909>

Slides: <https://purl.stanford.edu/tw704kw8870> DOI: <https://doi.org/10.25740/tw704kw8870>

Lecture 8 The quantum view of the world 4

Lecture 8a - The nature of quantum mechanical particles

MPES 3.6



This lecture part deals with the nature of quantum mechanical particles, or their "ontology", pointing out that many of the problems we have with understanding quantum mechanical particles come from us bringing along the ontology of classical particles when we use the word "particle" in quantum mechanics. Many of these problems disappear if we stop doing that, and there are classical analogies that can help with the concepts.

Keywords: Ontology and quantum mechanics, Quantum mechanical particles

Video: <https://purl.stanford.edu/hz252wh8692> DOI: <https://doi.org/10.25740/hz252wh8692>

Slides: <https://purl.stanford.edu/ny971sf7296> DOI: <https://doi.org/10.25740/ny971sf7296>

Lecture 8b - Waves and measurement MPES 3.6



This lecture part discusses the subtle issues of quantum mechanical waves and measurement. Such waves may not themselves be measurable entities, and that resolves some of the apparent contradictions of quantum mechanics. This lecture part extends the discussion of quantum mechanical measurement. It discusses Born's hypothesis that is essentially that the modulus squared of the wavefunction (not the wavefunction itself) gives probabilities in measurement. This hypothesis is also generalized in quantum mechanics to give probabilities of other quantities after measurement. A brief discussion is also given of the measurement problem more generally - the notion that we may not be able to describe the measurement process using quantum mechanics - and some of the proposed solutions to this problem.

Keywords: Measurement problem in quantum mechanics, Born's rule

Video: <https://purl.stanford.edu/mh418vy8682> DOI: <https://doi.org/10.25740/mh418vy8682>

Slides: <https://purl.stanford.edu/qz315cc5681> DOI: <https://doi.org/10.25740/qz315cc5681>

Lecture 9 The quantum view of the world 5

Lecture 9a - Tunneling MPES 3.7



This lecture part introduces the idea of quantum mechanical tunneling, in which a particle such as an electron can penetrate into or even through a barrier that is too "high". This is a core behavior in quantum mechanics and occurs routinely in electronic devices, for example. The problem of an electron and a finite barrier is solved to describe this behavior.

Keywords: Quantum mechanical tunneling

Video: <https://purl.stanford.edu/wj806yg0901> DOI: <https://doi.org/10.25740/wj806yg0901>

Slides: <https://purl.stanford.edu/xs652wr9105> DOI: <https://doi.org/10.25740/xs652wr9105>

Lecture 9b - Solving for barriers of finite height MPES 3.7



This lecture part gives a more detailed solution for an electron tunneling "into" a barrier, including the necessary "boundary conditions" in quantum mechanics for solving the Schroedinger equation in detail for this problem.

Keywords: Quantum mechanical tunneling

Video: <https://purl.stanford.edu/bt995tb8796> DOI: <https://doi.org/10.25740/bt995tb8796>

Slides: <https://purl.stanford.edu/dz392qh4115> DOI: <https://doi.org/10.25740/dz392qh4115>

Lecture 9c - Tunneling through a barrier MPES 3.7



This lecture part shows the solutions for a particle tunneling through a barrier that is too "high", showing the wave phenomena of both partial transmission and partial reflection.

Keywords: Quantum mechanical tunneling

Video: <https://purl.stanford.edu/zj462wx5618> DOI: <https://doi.org/10.25740/zj462wx5618>

Slides: <https://purl.stanford.edu/yf622br4239> DOI: <https://doi.org/10.25740/yf622br4239>

Lecture 10 Particles, atoms, and crystals 1

Lecture 10a - The hydrogen atom and center-of-mass coordinates MPES 4.1-4.3



This lecture part starts on the quantum mechanical solution of the hydrogen atom, a major triumph of quantum mechanics and the ultimate basis for much of our understanding of all atoms in chemistry. Here the "center of mass" coordinate view of the electron and the proton in the hydrogen atom is constructed as the first part of this solution.

Keywords: Hydrogen atom, Center of mass coordinates

Video: <https://purl.stanford.edu/zw467cn4933> DOI: <https://doi.org/10.25740/zw467cn4933>

Slides: <https://purl.stanford.edu/vd033ck8060> DOI: <https://doi.org/10.25740/vd033ck8060>

Lecture 10b - The hydrogen atom solutions and angular behavior MPES 4.3-4.4



This lecture part continues the quantum mechanical solution of the hydrogen atom, starting the discussion of the angular behavior of resulting wavefunction solution, and introducing the so-called spherical harmonics that ultimately lead to the shape of atomic orbitals.

Keywords: Spherical harmonics, Hydrogen atom

Video: <https://purl.stanford.edu/gm561gs5836> DOI: <https://doi.org/10.25740/gm561gs5836>

Slides: <https://purl.stanford.edu/rt925jm7079> DOI: <https://doi.org/10.25740/rt925jm7079>

Lecture 10c - Spherical harmonics for a classical problem MPES 4.4



This lecture part gives a simple visualization of the spherical harmonic functions that form part of the hydrogen atom solution. These functions are also the solutions to some classical problems, including the vibrations of a spherical shell. Visualizing those vibrations gives a useful and simple view of the nature of spherical harmonic functions.

Keywords: Spherical harmonics, Hydrogen atom

Video: <https://purl.stanford.edu/qb985kq9531> DOI: <https://doi.org/10.25740/qb985kq9531>

Slides: <https://purl.stanford.edu/sj577wq4370> DOI: <https://doi.org/10.25740/sj577wq4370>

Lecture 10d - Polar plots of spherical harmonics MPES 4.4



This lecture part shows spherical harmonics in so-called "polar" plots, which is the typical way they are shown in discussions of atomic orbitals in chemistry.

Keywords: Spherical harmonics, Hydrogen atom

Video: <https://purl.stanford.edu/jq656hz4970> DOI: <https://doi.org/10.25740/jq656hz4970>

Slides: <https://purl.stanford.edu/sz761sb9879> DOI: <https://doi.org/10.25740/sz761sb9879>

Lecture 10e - Spherical harmonics and atomic orbitals MPES 4.4



This lecture part relates the description of spherical harmonics in terms of quantum numbers (usually l and m in quantum mechanics) to the "spdf" notation common in discussing atomic orbitals in chemistry, and relates them to angular momentum in quantum mechanics.

Keywords: Spherical harmonics, Hydrogen atom, Angular momentum

Video: <https://purl.stanford.edu/rq562fh3064> DOI: <https://doi.org/10.25740/rq562fh3064>

Slides: <https://purl.stanford.edu/vr825xw4281> DOI: <https://doi.org/10.25740/vr825xw4281>

Lecture 11 Particles, atoms, and crystals 2

Lecture 11a - Hydrogen atom radial solutions MPES 4.5



This lecture part shows the radial parts of the solution for the hydrogen atom wavefunction

Keywords: Hydrogen atom, Schrödinger equation

Video: <https://purl.stanford.edu/sd796kz9548> DOI: <https://doi.org/10.25740/sd796kz9548>

Slides: <https://purl.stanford.edu/ks519mh9217> DOI: <https://doi.org/10.25740/ks519mh9217>

Lecture 11b - Hydrogen atom complete solutions MPES 4.5–4.6



This lecture part shows simulated images of the shape of the hydrogen atom wavefunctions (or, to be more precise, the probability densities) for various of these "orbitals". It also summarizes the meaning and use of the three major quantum numbers, n , l , and m , used to describe these wavefunctions or orbitals.

Keywords: Hydrogen atom, Schrödinger equation, Hydrogen orbitals

Video: <https://purl.stanford.edu/rb635yj1151> DOI: <https://doi.org/10.25740/rb635yj1151>

Slides: <https://purl.stanford.edu/dd805pk8481> DOI: <https://doi.org/10.25740/dd805pk8481>

Lecture 11c - Electron spin and Pauli exclusion MPES 4.7



This lecture part introduces the idea of electron spin. Together with the Pauli exclusion principle, which states that only one electron can occupy a given quantum mechanical state, these can explain the occupation of "orbitals" or atomic states in atoms.

Keywords: Quantum mechanical spin, Pauli exclusion principle

Video: <https://purl.stanford.edu/jb407pr1284> DOI: <https://doi.org/10.25740/jb407pr1284>

Slides: <https://purl.stanford.edu/yk088ct4957> DOI: <https://doi.org/10.25740/yk088ct4957>

Lecture 12 Particles, atoms, and crystals 3

Lecture 12a - Many-electron atoms MPES 4.8



This lecture part starts the discussion of atoms other than hydrogen, which are necessarily "many-electron" atoms. Many of the ideas that are exact for the hydrogen atom can be applied approximately to these atoms, including the idea that the potential energy is approximately still "central" (depending approximately only on the distance from the center of the nucleus and approximately not depending on angle).

Keywords: Many-electron atoms, Central potential

Video: <https://purl.stanford.edu/wz531gi0265> DOI: <https://doi.org/10.25740/wz531gi0265>

Slides: <https://purl.stanford.edu/qy836fd1894> DOI: <https://doi.org/10.25740/qy836fd1894>

Lecture 12b - Filling "shells" in atoms MPES 4.8



This lecture part introduces the idea of electrons filling successive "shells" in atoms. This idea helps in understanding chemical properties. A useful rule for the order of the filling of the shells is Madelung's rule. Though not always correct, it works for most atoms, and is a useful guide. Examples are given.

Keywords: Madelung's rule, Atomic shells, Atomic orbitals, Many-electron atoms

Video: <https://purl.stanford.edu/nd410hn2402> DOI: <https://doi.org/10.25740/nd410hn2402>

Slides: <https://purl.stanford.edu/yc280kj4124> DOI: <https://doi.org/10.25740/yc280kj4124>

Lecture 12c - Fermions and bosons MPES 4.9



This lecture part introduces the ideas of fermions, which have half integer spin (usually $1/2$), and bosons, which have integer spin (for example, 1). Fermions obey Pauli exclusion, but bosons do not. Common misconceptions are also discussed.

Keywords: Fermions, Bosons, Pauli exclusion principle

Video: <https://purl.stanford.edu/xh278hq3961> DOI: <https://doi.org/10.25740/xh278hq3961>

Slides: <https://purl.stanford.edu/fd833vr4122> DOI: <https://doi.org/10.25740/fd833vr4122>

Lecture 12d - Identical particles MPES 4.9



This lecture part describes the quantum mechanical notion of identical particles. Particles in quantum mechanics are identical in the way that dollars in a bank account are identical, in contrast to dollar bills, which are never identical (they have different serial numbers). This leads to quite different counting of possible states of multiple particles.

Keywords: Identical particles, Fermions, Bosons

Video: <https://purl.stanford.edu/cn674qm1512> DOI: <https://doi.org/10.25740/cn674qm1512>

Slides: <https://purl.stanford.edu/qf727hz2590> DOI: <https://doi.org/10.25740/qf727hz2590>

Lecture 13 Particles, atoms, and crystals 4

Lecture 13a - Coupled systems MPES 4.10



This lecture part introduces the idea of coupled systems in quantum mechanics, using the example of two "rectangular" potential wells with a thin barrier between them. What would have been isolated states in two separate wells turn into two coupled states, a "symmetric" one generally with lower energy and an "antisymmetric" one generally with higher energy. These lower and higher energy states introduce an idea common in understanding some kinds of chemical bonds, with these states representing "bonding" and "antibonding" states.

Keywords: Chemical bonding, Chemical antibonding, Coupled wells

Video: <https://purl.stanford.edu/jk870rb4220> DOI: <https://doi.org/10.25740/jk870rb4220>

Slides: <https://purl.stanford.edu/dn955nd5743> DOI: <https://doi.org/10.25740/dn955nd5743>

Lecture 13b - Crystals MPES 4.11



This lecture part introduces the idea of crystals - materials whose properties are periodic in space - together with key concepts like unit cells and crystal lattices. Example lattices in common electronic and optoelectronic devices, such as "zinc blende" and diamond lattices, are shown. The idea of semiconductor alloys - approximately crystalline materials made out of mixtures of atoms - is introduced because of their practical importance in devices. Crystalline growth is briefly discussed.

Keywords: Crystals, Crystals > Growth, Diamond lattice, Zinc blende lattice, Epitaxial growth

Video: <https://purl.stanford.edu/fb483zt1193> DOI: <https://doi.org/10.25740/fb483zt1193>

Slides: <https://purl.stanford.edu/tb447jx4506> DOI: <https://doi.org/10.25740/tb447jx4506>

Lecture 13c - Emergence of bands MPES 4.12



This lecture part illustrates how, when we bring N identical potential wells (or atoms) together, the individual well or atomic energy levels split into N different coupled energy levels, which we can view as forming a "band" of energies or states. These wavefunctions tend to be a product of a "unit cell" function that is approximately the same in each unit cell or period, and a larger envelope function, which is approximately a large sinusoid.

Keywords: Band structures

Video: <https://purl.stanford.edu/rg362bk0501> DOI: <https://doi.org/10.25740/rg362bk0501>

Slides: <https://purl.stanford.edu/fq775kk6226> DOI: <https://doi.org/10.25740/fq775kk6226>

Lecture 14 Particles, atoms, and crystals 5

Lecture 14a - Band structures in crystals MPES 4.13



This lecture part introduces the idea of band structures in crystals and the important "single electron" approximation of presuming that each electron moves in a periodic potential given by all the other electrons and nuclei in the crystal.

Keywords: Band structures, One-electron approximation

Video: <https://purl.stanford.edu/qy203yt0939> DOI: <https://doi.org/10.25740/qy203yt0939>

Slides: <https://purl.stanford.edu/dt345vm5726> DOI: <https://doi.org/10.25740/dt345vm5726>

Lecture 14b - The Bloch theorem MPES 4.13



This lecture part introduces the Bloch theorem, an important concept that allows us to think of wavefunctions in crystals as a product of a unit cell function that is the same in every unit cell of the crystal and an envelope function with a crystal wavevector. It gives the basic concept necessary for the idea of band structures in crystals.

Keywords: Bloch theorem, Band structures

Video: <https://purl.stanford.edu/mq967ch0040> DOI: <https://doi.org/10.25740/mq967ch0040>

Slides: <https://purl.stanford.edu/mx164qy5831> DOI: <https://doi.org/10.25740/mx164qy5831>

Lecture 14c - Solving with periodic boundary conditions MPES 4.13



This lecture part explains the idea of periodic boundary conditions, which is technically an approximation but which works well in large crystals, and allows the Bloch theorem. This idea also leads an alternate form for the Bloch theorem. An example simple Bloch form wavefunction is shown.

Keywords: Bloch theorem, Periodic boundary conditions, Crystals

Video: <https://purl.stanford.edu/cb769pm7614> DOI: <https://doi.org/10.25740/cb769pm7614>

Slides: <https://purl.stanford.edu/jd092yb8124> DOI: <https://doi.org/10.25740/jd092yb8124>

Lecture 15 Particles, atoms, and crystals 6

Lecture 15a - Band structures MPES 4.14



This lecture part illustrates how a simple band structure is built up by solving for the energies at each of a set of k (wavevector) values. It shows the concept of a Brillouin zone - the range of k required to represent a band structure - as well as concepts like a band gap energy between two bands, and the extended zone scheme, which shows the band structure just repeats for other values of k .

Keywords: Band structure, Brillouin zone, Energy band gap

Video: <https://purl.stanford.edu/hm072cj4906> DOI: <https://doi.org/10.25740/hm072cj4906>

Slides: <https://purl.stanford.edu/yk259md0718> DOI: <https://doi.org/10.25740/yk259md0718>

Lecture 15b - Crystal momentum and effective mass MPES 4.14



This lecture part introduces the ideas of crystal momentum, the effective momentum of the electron in the periodic potential, and effective mass, the apparent mass of the electron because of the band structure. It also introduces the ideas of indirect gap and direct gap materials.

Keywords: Effective mass, Crystal momentum, Direct gap, Indirect gap

Video: <https://purl.stanford.edu/nq967fr6621> DOI: <https://doi.org/10.25740/nq967fr6621>

Slides: <https://purl.stanford.edu/vy907jq6338> DOI: <https://doi.org/10.25740/vy907jq6338>

Lecture 15c - Band structures in three dimensions MPES 4.15



This lecture part formally extends the ideas of band structures and Brillouin zones from the one-dimensional illustrations to three dimensions, as in most actual crystals

Keywords: Band structures, Brillouin zones, Crystals

Video: <https://purl.stanford.edu/tc664jg4027> DOI: <https://doi.org/10.25740/tc664jg4027>

Slides: <https://purl.stanford.edu/vw087nj3876> DOI: <https://doi.org/10.25740/vw087nj3876>

Lecture 15d - Plotting actual band structures MPES 4.15



The lecture part shows how band structures are plotted in practice for real materials, such as silicon and "III-V" (3-5) semiconductors such as Gallium Arsenide, including the concepts of conduction bands and valence bands that are important for devices.

Keywords: Band structures, Silicon, Gallium Arsenide, Conduction band, Valence band, Crystals

Video: <https://purl.stanford.edu/x935zh6734> DOI: <https://doi.org/10.25740/x935zh6734>

Slides: <https://purl.stanford.edu/fd348by1209> DOI: <https://doi.org/10.25740/fd348by1209>

Lecture 15e - Metals, semiconductors and insulators MPES 4.16



This lecture part shows how semiconductors, insulators, and metals are distinguished by the different forms of band structures and by the size of bandgap energies.

Keywords: Semiconductors, Metals, Insulators, Conduction band, Valence band, Electrons and holes

Video: <https://purl.stanford.edu/cd733wr3720> DOI: <https://doi.org/10.25740/cd733wr3720>

Slides: <https://purl.stanford.edu/pc321yn1921> DOI: <https://doi.org/10.25740/pc321yn1921>

Lecture 16 Thermal distributions 1

Lecture 16a - Tossing coins, microstates, and macrostates MPES 5.2



This lecture part starts the discussion of statistical mechanics by looking at the outcomes of tossing coins, including the resulting distributions of "heads" and "tails". It also introduces the concepts of microstates, macrostates, and multiplicity.

Keywords: Statistical mechanics, Microstates, Macrostates, Multiplicity

Video: <https://purl.stanford.edu/fb753nr5592> DOI: <https://doi.org/10.25740/fb753nr5592>

Slides: <https://purl.stanford.edu/pz940zh4180> DOI: <https://doi.org/10.25740/pz940zh4180>

Lecture 16b - Binomial distribution and Stirling's approximation MPES 5.2



This lecture part introduces the binomial distribution as the formal description useful for discussing distributions of "heads" and "tails", for example. It then introduces Stirling's approximation, which is very useful for large numbers of particles (or coin tosses), and leads to simple formulas. This also gives simple results for how wide we expect the distribution (of multiplicities) to be.

Keywords: Binomial distribution, Stirling's approximation, Statistical mechanics

Video: <https://purl.stanford.edu/tt799kx6395> DOI: <https://doi.org/10.25740/tt799kx6395>

Slides: <https://purl.stanford.edu/tj109rj4343> DOI: <https://doi.org/10.25740/tj109rj4343>

Lecture 16c - Two-state spin systems and microstates MPES 5.3



This lecture part moves to discussing actual physical systems, here of two-state spin systems (such as electrons), and adds the possibility that the "spin up" and "spin down" states could have different energies. This then gives a model system for introducing many physical concepts and behaviors. It also introduces the concept of accessible microstates and basic assumptions often used in statistical mechanics.

Keywords: Electron spin, Statistical mechanics

Video: <https://purl.stanford.edu/sz568xk7258> DOI: <https://doi.org/10.25740/sz568xk7258>

Slides: <https://purl.stanford.edu/ct742rm1668> DOI: <https://doi.org/10.25740/ct742rm1668>

Lecture 17 Thermal distributions 2

Lecture 17a - Systems in thermal contact MPES 5.3



This lecture part introduces the idea of putting two systems of spins in thermal contact, which means they can exchange energy through some thermally conducting wall, and examines the multiplicities of the combined system.

Keywords: Statistical mechanics, Thermal conduction

Video: <https://purl.stanford.edu/wh168wn2090> DOI: <https://doi.org/10.25740/wh168wn2090>

Slides: <https://purl.stanford.edu/xy900nx7103> DOI: <https://doi.org/10.25740/xy900nx7103>

Lecture 17b - Maximizing multiplicities for spin systems MPES 5.3



This lecture part examines what happens if we maximize the multiplicity of two spin systems joined by a thermally conducting wall. This should correspond to the most likely situation, with the result that the energy per spin becomes the same on both sides of the wall.

Keywords: Statistical mechanics, Thermal equilibrium

Video: <https://purl.stanford.edu/bn960qk9816> DOI: <https://doi.org/10.25740/bn960qk9816>

Slides: <https://purl.stanford.edu/fw715fd7177> DOI: <https://doi.org/10.25740/fw715fd7177>

Lecture 17c - Maximizing multiplicity for general systems MPES 5.4



This lecture part generalizes the idea of maximizing multiplicity by thermal conduction, with a simple result for the quantity that becomes the same on both sides.

Keywords: Statistical mechanics, Thermal equilibrium

Video: <https://purl.stanford.edu/pf498qv1024> DOI: <https://doi.org/10.25740/pf498qv1024>

Slides: <https://purl.stanford.edu/gk207pc8527> DOI: <https://doi.org/10.25740/gk207pc8527>

Lecture 18 Thermal distributions 3

Lecture 18a - Entropy and temperature MPES 5.4



In this lecture part, entropy is defined in terms of the logarithm of the multiplicity. Based on what maximizes the multiplicity in thermal equilibration, and hence what maximizes entropy, in turn that leads to a definition of temperature in terms of entropy. An example system is shown.

Keywords: Statistical mechanics, Entropy, Temperature

Video: <https://purl.stanford.edu/dz224jd9442> DOI: <https://doi.org/10.25740/dz224jd9442>

Slides: <https://purl.stanford.edu/vn915mc5638> DOI: <https://doi.org/10.25740/vn915mc5638>

Lecture 18b - Entropy and heat flow MPES 5.4



This lecture part describes how the flow of heat from hotter to colder bodies leads to an increase in entropy overall, with the entropy of the hotter body decreasing but the entropy of the cooler body increasing by more. A numerical example is given to show just how enormously unlikely it is that even a small amount of heat would instead flow from cold to hot. In turn, this justifies the second law of thermodynamics.

Keywords: Statistical mechanics, Second law of thermodynamics, Entropy, Temperature

Video: <https://purl.stanford.edu/zx615zc5253> DOI: <https://doi.org/10.25740/zx615zc5253>

Slides: <https://purl.stanford.edu/mh342xb8835> DOI: <https://doi.org/10.25740/mh342xb8835>

Lecture 18c - Carnot efficiency limit for heat engines MPES 5.5



Based on the idea that entropy cannot decrease for any large (closed) system, this lecture part derives the so-called "Carnot" limit to the efficiency of any kind of heat engine (examples include steam engines, refrigerators, air conditioners, internal combustion engines, heat pumps).

Keywords: Carnot efficiency, Entropy, Second law of thermodynamics

Video: <https://purl.stanford.edu/yn917qh7340> DOI: <https://doi.org/10.25740/yn917qh7340>

Slides: <https://purl.stanford.edu/kb702dz3702> DOI: <https://doi.org/10.25740/kb702dz3702>

Lecture 19 Thermal distributions 4

Lecture 19a - The Boltzmann factor MPES 5.6



This lecture part introduces the Boltzmann factor, which gives the relative probability of two states of different energies at some temperature.

Keywords: Statistical mechanics, Boltzmann factor

Video: <https://purl.stanford.edu/xw084gh9764> DOI: <https://doi.org/10.25740/xw084gh9764>

Slides: <https://purl.stanford.edu/ym134ms6380> DOI: <https://doi.org/10.25740/ym134ms6380>

Lecture 19b - Chemical potential MPES 5.7



This lecture part introduces the chemical potential, which is the quantity that is equalized when particles are allowed to diffuse through a wall.

Keywords: Chemical potential, Statistical mechanics

Video: <https://purl.stanford.edu/xh451ny4159> DOI: <https://doi.org/10.25740/xh451ny4159>

Slides: <https://purl.stanford.edu/xx959cm6376> DOI: <https://doi.org/10.25740/xx959cm6376>

Lecture 19c - Chemical potential and the Gibbs factor MPES 5.7



This lecture part introduces the Gibbs factor, which gives the relative probability of two states with possibly different energies and different numbers of particles.

Keywords: Statistical mechanics, Gibbs factor

Video: <https://purl.stanford.edu/wt576sj3412> DOI: <https://doi.org/10.25740/wt576sj3412>

Slides: <https://purl.stanford.edu/vp977ch0604> DOI: <https://doi.org/10.25740/vp977ch0604>

Lecture 20 Thermal distributions 5

Lecture 20a - The Fermi-Dirac distribution MPES 5.8



Using the Gibbs factor (and typically calling the chemical potential the Fermi energy instead), this lecture part derives the Fermi-Dirac distribution, which gives the occupation probability of states for fermions (such as electrons) as a function of temperature and Fermi energy.

Keywords: Fermi-Dirac distribution, Chemical potential, Fermi energy

Video: <https://purl.stanford.edu/nz615sh8066> DOI: <https://doi.org/10.25740/nz615sh8066>

Slides: <https://purl.stanford.edu/cc251qv7462> DOI: <https://doi.org/10.25740/cc251qv7462>

Lecture 20b - The Bose-Einstein and Planck distributions MPES 5.8



This lecture part states the Bose-Einstein distribution, which gives the number of bosons that should be expected on average in any given state at a given temperature, and derives the simpler version that is the Planck distribution, which applies to photons in particular, telling us the number of photons expected in a photon mode in thermal equilibrium.

Keywords: Bose-Einstein distribution, Planck distribution, Statistical mechanics

Video: <https://purl.stanford.edu/tw706mh0577> DOI: <https://doi.org/10.25740/tw706mh0577>

Slides: <https://purl.stanford.edu/kb974yz8815> DOI: <https://doi.org/10.25740/kb974yz8815>

Lecture 20c - The Maxwell-Boltzmann distribution MPES 5.8



This lecture part states the Maxwell-Boltzmann distribution, which would be the thermal distribution of non-identical particles. Though quantum mechanical particles of a given kind are actually identical, this distribution is often used in practice in some situations because it gives a simpler mathematical form that is valid anyway as the high-temperature high-energy "tail" of both the Fermi-Dirac and Bose-Einstein distribution.

Keywords: Maxwell-Boltzmann distribution, Statistical mechanics

Video: <https://purl.stanford.edu/zs242pm5949> DOI: <https://doi.org/10.25740/zs242pm5949>

Slides: <https://purl.stanford.edu/jh484rj8773> DOI: <https://doi.org/10.25740/jh484rj8773>

Lecture 21 Bands and electronic devices

Lecture 21a - Electrons and holes in bands MPES 6.2



This lecture part introduces the concept of how electrons and holes move or "transport" in materials, including the "drift" model that leads to conventional electrical resistance. It also expands on the idea of holes and how we view their energy levels.

Keywords: Electrons and holes, Drift transport, Semiconductors

Video: <https://purl.stanford.edu/bp965yk4656> DOI: <https://doi.org/10.25740/bp965yk4656>

Slides: <https://purl.stanford.edu/sz835sz0768> DOI: <https://doi.org/10.25740/sz835sz0768>

Lecture 21b - Semiconductor doping and diodes MPES 6.3, 6.4



This lecture part discusses n-type and p-type doping in semiconductors and the resulting diodes that can be formed in this way. It also discusses that Fermi levels are equalized when such materials are joined together, because of the movement of electrons and holes to equalize their chemical potentials (Fermi levels), and shows the resulting band diagrams for diodes.

Keywords: Semiconductor diodes, n-type doping, p-type doping, Fermi levels

Video: <https://purl.stanford.edu/tc688np4377> DOI: <https://doi.org/10.25740/tc688np4377>

Slides: <https://purl.stanford.edu/bx769xf8457> DOI: <https://doi.org/10.25740/bx769xf8457>

Lecture 21c - Voltages and Fermi levels MPES 6.4 (and 5.7 for equivalent chemical potential definition)



This lecture part shows why it is that, when we apply a voltage to a diode, we separate the Fermi levels on the two sides by an amount equal (in electron-volts) to the applied ("bias") voltage. This will allow us to understand biasing of diodes.

Keywords: Fermi levels, Semiconductor diodes

Video: <https://purl.stanford.edu/dg178rw7394> DOI: <https://doi.org/10.25740/dg178rw7394>

Slides: <https://purl.stanford.edu/tv852qw4189> DOI: <https://doi.org/10.25740/tv852qw4189>

Lecture 21d - Biasing semiconductor devices MPES 6.4



This lecture part shows what happens when we apply "bias" voltages to diodes. Then, using the statistical mechanics of electrons and holes in the Maxwell-Boltzmann approximation, we are able to derive the ideal diode current-voltage characteristic, giving a good example of putting together many different ideas.

Keywords: Semiconductor diodes

Video: <https://purl.stanford.edu/nd351qg2808> DOI: <https://doi.org/10.25740/nd351qg2808>

Slides: <https://purl.stanford.edu/sv627gw7791> DOI: <https://doi.org/10.25740/sv627gw7791>

Lecture 22 Light and quantum mechanics 1

Lecture 22a - Light, the photoelectric effect, and the photon MPES 7.2



This lecture part describes the photoelectric effect and how Einstein's proposal of the photon led to its explanation.

Keywords: Photoelectric effect

Video: <https://purl.stanford.edu/ww801fb8540> DOI: <https://doi.org/10.25740/ww801fb8540>

Slides: <https://purl.stanford.edu/zb916cm8302> DOI: <https://doi.org/10.25740/zb916cm8302>

Lecture 22b - Light and modes MPES 7.3



This lecture part introduces one way of understanding how many "modes" of light there are, modes that can be occupied by photons.

Keywords: Quantum mechanics of light, Optical modes

Video: <https://purl.stanford.edu/hm042vn5425> DOI: <https://doi.org/10.25740/hm042vn5425>

Slides: <https://purl.stanford.edu/qn632qm6607> DOI: <https://doi.org/10.25740/qn632qm6607>

Lecture 22c - Thermal radiation MPES 7.4



This lecture part derives Planck's description of the black-body radiation spectrum (as seen in the sun and in incandescent light bulbs, for example) that triggered much of modern quantum mechanics. It continues to derive the Stefan-Boltzmann law for the amount of energy in a light field in thermal equilibrium, which also explains why hot objects radiate so much light.

Keywords: Black body radiation, Stefan-Boltzmann law

Video: <https://purl.stanford.edu/rr796xy3214> DOI: <https://doi.org/10.25740/rr796xy3214>

Slides: <https://purl.stanford.edu/yk746kv0769> DOI: <https://doi.org/10.25740/yk746kv0769>

Lecture 23 Light and quantum mechanics 2

Lecture 23a - Black body radiation and Kirchhoff's law MPES 7.5



This lecture part shows how to envisage a perfect black body for thought experiments and derivations, and goes on to derive Kirchhoff's radiation law, which states that the "absorptivity" of an object (the fraction of incident light it absorbs) must equal its emissivity (the amount of light it emits at a given temperature as a fraction of the light a perfect black body would emit).

Keywords: Kirchhoff radiation law, Black body radiation

Video: <https://purl.stanford.edu/dy762sz3700> DOI: <https://doi.org/10.25740/dy762sz3700>

Slides: <https://purl.stanford.edu/cf521rq1709> DOI: <https://doi.org/10.25740/cf521rq1709>

Lecture 23b - Einstein's A and B coefficient argument MPES 7.6



This lecture part derives Einstein's "A & B coefficient" argument, which predicts the existence and strength of stimulated emission, the kind of emission we get from lasers.

Keywords: Einstein A and B coefficient argument, Lasers

Video: <https://purl.stanford.edu/pm246kg9828> DOI: <https://doi.org/10.25740/pm246kg9828>

Slides: <https://purl.stanford.edu/jg412dd7396> DOI: <https://doi.org/10.25740/jg412dd7396>

Lecture 24 Semiconductor optoelectronics 1

Lecture 24a - Optoelectronic devices and photodetectors MPES 8.1, 8.2



This lecture part introduces semiconductor optoelectronic devices, the devices that convert light to electricity and electricity to light. Devices include photoconductors, photodiodes.

Keywords: Semiconductor optoelectronic devices, Photoconductors, Photodiodes

Video: <https://purl.stanford.edu/hg132hn9284> DOI: <https://doi.org/10.25740/hg132hn9284>

Slides: <https://purl.stanford.edu/gp544zn8910> DOI: <https://doi.org/10.25740/gp544zn8910>

Lecture 24b - Light emission MPES 8.3



This lecture part introduces both thermal (e.g., incandescent light bulb) and non-thermal (e.g., light-emitting diodes and lasers) light sources. It goes on to consider the light emission from semiconductors, especially the strong emission from direct-gap semiconductors, and explains the light emission from semiconductor light-emitting diodes (LEDs).

Keywords: LEDs (Light emitting diodes), Direct gap semiconductors, Light emission

Video: <https://purl.stanford.edu/vf654xr7507> DOI: <https://doi.org/10.25740/vf654xr7507>

Slides: <https://purl.stanford.edu/sm396gq4063> DOI: <https://doi.org/10.25740/sm396gq4063>

Lecture 25 Semiconductor optoelectronics 2

Lecture 25a - Absorption and emission processes MPES 8.3



This lecture part introduces the absorption and emission processes, both spontaneous and stimulated, from a quantum mechanical point of view of transitions between energy levels.

Keywords: Optical absorption, Spontaneous emission, Stimulated emission

Video: <https://purl.stanford.edu/yg356fd8716> DOI: <https://doi.org/10.25740/yg356fd8716>

Slides: <https://purl.stanford.edu/vf719nt6353> DOI: <https://doi.org/10.25740/vf719nt6353>

Lecture 25b - Lasers MPES 8.3



This lecture part introduces how lasers work, including the ideas of population inversion and 3-level and 4-level laser systems, and semiconductor lasers in particular.

Keywords: Lasers, Semiconductor lasers, Population inversion

Video: <https://purl.stanford.edu/sn066wc4638> DOI: <https://doi.org/10.25740/sn066wc4638>

Slides: <https://purl.stanford.edu/mp376pd2384> DOI: <https://doi.org/10.25740/mp376pd2384>

Lecture 25c - Epilogue MPES Chapter 9



This lecture part is a short epilogue to the entire course of Modern Physics for Engineers and Scientists.

Keywords: Modern physics

Video: <https://purl.stanford.edu/tf597tt5819> DOI: <https://doi.org/10.25740/tf597tt5819>

Slides: <https://purl.stanford.edu/gf324tw8036> DOI: <https://doi.org/10.25740/gf324tw8036>

Index

A

absolute zero, 9, 199
absorptivity, 265
acceptor atom. *See* atom, acceptor
adiabatic, 210
aeolipile, 7
age of the universe, 204
alchemy, 2–3
al-Din, Taqi, 7
Alhazen, 2, 5
alloy, 151
 quaternary, 151
 ternary, 151
amorphous material, 152
Ampère, André-Marie, 6
Ångström, Anders, 58
angular frequency, 23
angular momentum, 63, 66, 67, 94, 121, 125–26, 134, 135
angular momentum quantum number. *See* orbital quantum number
antibonding state, 146, 172, 173, 181
anti-particle, 135
Arago, François, 6
Aristotle, 2, 3
associated Laguerre polynomials, 128
associated Legendre functions, 120, 121
atom, 2, 3, 19, 58, 62, 81, 85, 113, 140, 146, 172, 181, 225
 absorption and emission of photons, 266–69, 279, 286
 acceptor, 178, 180
 Bohr model, 63–67, 93, 121
 donor, 178, 180, 306, 307
 helium. *See* helium
 hydrogen. *See* hydrogen
 many-electron, 68, 139, 135–39, 158
atomic configuration, 138
atomic number, 135, 136, 138
Aufbau principle, 138
Avicenna, 2
azimuthal angle, 118
azimuthal plane, 121

B

Bacon, Francis, 3
ballistic transport, 237, 239
Balmer series, 58, 59, 65
Balmer, Johann, 58
band
 emergence, 153–57, 181
 isotropic parabolic, 235
 parabolic, 168
band gap, 167, 173, 176, 178, 180, 273, 274, 301, 303, 304, 305, 306, 307
 direct, 169, 174, 279–81
 indirect, 169, 174, 280, 281
band structure, 168, 169, 175, 176, 235, 238, 240, 245, 297, 299
basis
 function, 19, 50–52
 set, 46–47, 51–53
 state, 19
binary compound, 151
binomial
 coefficient, 185
 distribution, 185, 186
black body, 59–61, 262–65
 radiation, 262–65
 spectrum, 59–63, 63, 254, 259–62, 271
Black, Joseph, 8
Bloch theorem, 157, 158, 159–63, 170, 174, 181, 297
Bohr magneton, 293
Bohr model. *See* atom, Bohr model
Bohr radius, 65, 128
Bohr, Neils, 63, 64, 65, 66, 125, 129
Boltzmann factor, 212–15, 219, 220, 227, 228, 233, 261, 266, 267, 301
Boltzmann, Ludwig, 9, 58, 182, 200
Boltzmann's constant, 198, 199, 202, 204, 292
bonding state, 146, 172, 173, 181
Born postulate, 73, 96
Born rule. *See* Born postulate
Born, Max, 67, 68, 73
Bose, Satyendra Nath, 141
Bose-Einstein distribution, 142, 223, 224–28, 228, 229

boson, 139–42, 142, 228, 229
 boundary conditions
 barrier of finite height, 102–3, 106
 hard wall, 36, 75, 160
 infinitely high walls, 75
 periodic, 159–62
 Boyle, Robert, 3
 Brahe, Tycho, 4
 bra-ket notation. *See* Dirac notation
 Branca, Giovanni, 7
 breathing mode, 122
 Brillouin zone, 167, 168, 171–72, 173, 174, 237, 280, 281
 bucky ball, 150

C

caloric, 8
 carbon nanotube, 150
 Carnot cycle, 8, 210
 Carnot limit, 210, 211
 Carnot, Nicolas Léonard Sadi, 8, 210
 carrier, 302, 303
 carrier concentration, 299, 303
 intrinsic, 305
 Celsius, Anders, 199
 center-of-mass coordinates, 114–16
 central potential, 119, 136
 chemical potential, 215–21, 217, 218, 225, 228, 245, 246, 282
 Fermi level. *See* Fermi level
 Gibb's definition, 217
 chemical symbol, 138
 classical wave equation, 26–32, 38
 Clausius, Rudolf, 8
 completeness, 46–47, 50, 51, 52, 53, 57
 conduction band, 169, 172, 173, 174, 234–35, 235–39, 235, 243, 245, 247, 248, 250, 273, 279–81, 282, 286, 285–87, 287, 297, 300, 301, 302, 303
 Copernicus, Nicolaus, 3
 corpuscular theory of light, 5
 Coulomb attraction, 63, 117, 306
 Coulomb potential, 117
 Coulomb repulsion, 136
 Coulomb, Charles, 6
 coupled oscillator, 39–44
 coupled well, 143–48
 coupled wells
 multiple, 153–57
 covalent bonding, 146
 crystal, 148–53
 lattice, 149
 crystal momentum. *See* momentum, crystal

D

Dalton, John, 3
 de Broglie, Louis, 67
 de Broglie's hypothesis, 67–69, 70, 85, 87, 167
 degenerate doping, 242
 degenerate statistics, 300, 307
 Democritus, 2
 density of states, 240, 258, 259, 297, 299, 303
 effective. *See* effective density of states
 depletion region, 273–78
 Descartes, René, 3
 diamond lattice, 149, 150, 151, 171
 diesel engine, 209
 diffraction, 5, 6, 85–88
 angle, 86
 two-slit. *See* two-slit experiment
 diffusion, 243–45, 273
 diffusion current, 248–53, 275, 276
 diffusive equilibrium. *See* equilibrium, diffusive
 diode, 148, 239–53, 272–73, 289
 current-voltage characteristic, 250, 251
 laser. *See* laser diode
 light-emitting. *See* light-emitting diode
 photo. *See* photodiode
 photovoltaic, 275
 Dirac notation, 48
 Dirac, Paul, 48, 67, 68, 135
 Dirac's equation, 135
 direct transition. *See* vertical transition
 donor atom. *See* atom, donor
 donor binding energy, 306
 doping, 178, 239, 305, 306, 235–39, 308
 n-type, 178–79
 p-type, 178–79
 drift transport, 237–38, 239, 273
 drift velocity, 238

E

Edison, Thomas, 60
 effective density of states
 at the conduction band edge, 304
 at the valence band edge, 304
 effective mass, 158, 168–69, 235, 238, 287, 297, 298, 303, 304, 305, 306, 307
 eigen equation, 24
 eigenfunction, 12, 19, 24–26, 37, 45, 46, 50–57, 74–81, 96, 159, 162, 299
 eigenstate, 19, 78–79, 80, 94, 96, 141, 167, 290
 eigenvalue, 24, 34, 37, 42, 50, 51, 53, 57
 degenerate, 24, 133, 171, 299
 energy, 78–79, 143, 167

Einstein, Albert, 4, 62, 223, 226, 254, 256, 266, 268, 271
 Einstein's A and B coefficient argument, 62, 266–71, 279, 280, 284
 electric constant, 292
 electrical breakdown, 177
 electron, 63–67, 85, 92, 93, 112, 114, 178
 center of mass behavior, 114–16
 charge, 64, 116, 292
 core, 137, 172
 diffusion, 244, 249
 emission, 255–56
 Fermi level, 282
 Fermi-Dirac distribution, 221–24, 287, 299
 g factor, 292
 injection, 272, 279, 281
 mass, 81
 microscope, 85
 orbital, 117
 rest mass, 115, 292
 screening, 137
 shells, 136–37, 179
 spin, 68, 134, 141
 sweep-out, 275
 transport, 235–38
 valence, 175
 electron-hole pair, 273
 electron-positron pair, 141
 electron-volt, 64
 elementary charge. *See* electron:charge
 emissivity, 265
 Empedocles, 2, 5
 enthalpy, 218
 entropy, 8–9, 9, 58, 182, 198–212, 213, 220, 232, 246, 265, 290, 291, 301
 fundamental, 198, 200, 204
 thermodynamic, 9, 200
 epitaxial growth, 152
 equilibrium, 191, 205, 225, 227, 245, 302, 305
 diffusive, 216, 217, 245, 247, 249, 282
 thermal, 192, 194, 197, 198, 214, 247, 259–68, 270, 279, 282, 285, 299
 Euclid, 5

F

face-centered cubic lattice, 150, 151, 171
 Fahrenheit, Daniel, 8
 Faraday, Michael, 7
 Fermi energy. *See* Fermi level
 Fermi level, 219, 223, 221–24, 225, 239, 243, 245, 247, 248, 249, 250, 256, 282, 286, 287, 299, 300, 302, 303, 306, 307
 quasi. *See* quasi Fermi level

 relation to chemical potential, 223
 temperature-dependence, 224
 Fermi, Enrico, 141
 Fermi-Dirac distribution, 142, 221–24, 225, 228, 229, 239, 256, 287, 298, 299, 301, 302, 303, 304, 307
 fermion, 139–42, 142, 221, 222, 225, 228, 229
 fine structure constant, 292, 293
 first law of thermodynamics. *See* thermodynamics, first law of
 Fizeau, Hippolyte, 6
 fluorescence, 5
 fluorescent light bulb, 272, 279
 force, 14, 15–17, 21, 28–30, 41, 51, 73, 116–18, 117, 125, 168, 236–37, 238, 255, 293, 306
 Franklin, Benjamin, 6
 Fresnel, Augustin-Jean, 6
 functional analysis, 51

G

Galileo, 3, 4, 7
 gambler's fallacy, 184
 gas turbine, 209
 gasoline engine, 209
 Gaussian, 184, 187, 232
 Gibbs factor, 215–21, 221, 222, 224, 233, 301
 Gibbs free energy, 218
 Gibbs, Josiah, 9, 182
 Gilbert, William, 6
 glass, 152
 gradient, 15
 operator, 16
 graphene, 150
 graphite, 150
 graviton, 141
 Grimaldi, Francesco, 5
 group velocity, 176, 236, 238

H

Hamilton, William, 4
 harmonics, 20
 Hartree-Fock method, 136
 heat engine, 7, 8, 9, 10, 205, 209, 211
 heat flow, 201–5
 Heisenberg matrix model, 67, 71, 112
 Heisenberg, Werner, 67
 Heisenberg's uncertainty principle, 88, 85–89
 helium, 141
 Helmholtz free energy, 218
 Helmholtz resonator, 20
 Helmholtz wave equation, 32–36, 36, 69–71
 Hermitian adjoint, 47–50

Hermitian matrix, 50
 Hermitian operator, 50–51, 52–57, 57
 Hero of Alexandria, 7
 Hertz, Heinrich, 58, 255
 Higgs boson, 141
 hole, 179, 180, 248, 273, 274, 280, 281, 286, 287, 302, 304
 diffusion, 244, 249, 250
 Fermi level, 282
 Fermi-Dirac distribution, 287, 304
 injection, 272, 279, 281
 sweep-out, 275
 transport, 238–39
 HOMO, 173
 Huygens Christiaan, 5
 Huygens' principle, 6, 89
 hydrogen, 58–59, 67, 63–69, 71–73, 79, 94, 112, 134, 135–37, 143, 163, 181
 orbital. *See* orbital, hydrogen atom
 radial wavefunction solutions, 126–32

I

Ibn Sahl, 5
 identical particles, 142–43, 228
 indirect transition, 281
 infinitely deep potential well, 74
 inner product, 44–45, 48
 insulator, 152, 173, 234–35, 301, 303, 307
 intrinsic, 240
 intrinsic semiconductor, 243
 inverted population, 287, *See* population inversion
 isothermal, 210
 isotropic parabolic band, 296, 297, 298

J

Jabir, 2
 James Watt, 7
 Jeans, James, 60
 jet engine, 209
 Joule, James, 8

K

Kelvin, Lord. *See* Thomson, William
 Kepler, Johannes, 4
 Kepler's laws, 4
 kinetic energy, 13, 80, 99, 146, 239, 255, 256
 Kirchhoff, Gustav, 60
 Kirchhoff's law of radiation, 60, 254, 264–65
 Kramers degeneracy, 171
 Kuo, Shen, 6

L

Lagrange, Joseph-Louis, 4
 laser, 1, 254, 266, 270, 271, 272, 282–87
 four-level, 286
 gain, 285
 three-level, 286
 laser diode, 283, 285–87, 289
 latent heat, 8
 Lavoisier, Antoine, 3, 8
 law of averages, 185
 law of large numbers, 185
 law of mass action, 303
 LED. *See* light-emitting diode
 Lenard, Philipp, 255
 Leucippus, 2
 Leyden jar, 6
 light-emitting diode, 1, 81, 150, 152, 279–82, 289
 logarithm, 186
 luminescence, 5, 284
 LUMO, 173

M

macrostate, 183, 184, 185, 188, 192, 193, 196, 197, 198, 202, 204, 205, 232, 290
 Madelung's rule, 137–39
 magnetic compass, 6
 magnetic constant, 292
 magnetic moment, 134, 191
 magnetic quantum number, 67, 121, 125, 126, 134
 magnetic spin quantum number, 134
 majority carrier, 248–53
 Maxwell, James Clerk, 7, 9, 58, 182
 Maxwell's equations, 1, 7, 62
 Maxwell-Boltzmann approximation. *See* Maxwell-Boltzmann distribution
 Maxwell-Boltzmann distribution, 142, 223, 228–32, 228, 229, 250, 303, 304, 306, 307
 measurement problem, 89, 95–99, 113
 Mendeleev, Dmitri, 3
 metal, 58, 59, 152, 234–35, 255, 256, 300
 microstate, 183, 184, 188, 190, 191, 192, 193, 196, 200, 201, 202, 204, 205, 212, 213, 219, 220, 232
 minority carrier, 248–53
 mobility, 238
 mode, 12
 boson, 224, 225, 226, 229
 definition, 52
 electromagnetic, 94, 141, 259–62, 256–60, 269, 270, 283, 284
 fundamental, 20
 light, 256–60

normal, 19
 oscillating, 23, 53, 122
 propagating, 20, 94
 spherical shell, 122–24
 standing wave, 36–39, 94
 momentum, 237
 angular. *See* angular momentum
 classical, 4, 13, 15, 93, 94
 crystal, 167–69, 235, 237, 281
 photon, 87, 280
 pseudo, 168
 quantum mechanical, 67–69, 85–89, 94
 monochromatic. *See* wave, monochromatic
 multiplicity, 183, 184, 185, 186, 187, 192, 193,
 196, 197, 198, 200, 205, 213, 220, 232, 291

N

natural logarithm. *See* logarithm
 neutron spin, 141
 Newcomen, Thomas, 7
 Newton, Isaac, 4, 5
 Newton's laws of motion, 1, 4–5, 15, 21, 30, 41
 nodal circles, 122–24
 non-degenerate statistics, 299, 307
 normalization, 73, 76–78, 95
*N*th roots of unity. *See* roots of unity

O

Ohm's law, 238
 one-electron approximation, 158–59
 ontology, 93
 operator
 definition, 24
 eigenfunction, 52–57
 linear, 25–26, 34, 42
 optoelectronic device, 272
 orbital, 19, 67, 94, 116, 117, 119, 150, 172, 173,
 181, 221
 hydrogen atom, 125–33
 many-electron atom, 135–39
 notation, 126
 orbital quantum number, 67, 121, 126, 134
 Orsted (Ørsted), Hans Christian, 6
 orthogonality, 20, 44–47, 50–51, 53, 57, 80, 133

P

Parmenides, 2
 particle-in-a-box model, 73–85, 94, 113, 114, 143,
 156
 partition function, 222, 224, 227

Pauli exclusion principle, 134–35, 141, 181, 221,
 229
 Pauli, Wolfgang, 67, 68, 71, 134, 135
 periodic potential, 158
 periodic table, 3, 9, 58, 150, 179
 permeability
 of free space. *See* magnetic constant
 permittivity
 of free space. *See* electric constant
 phase, 23
 phase velocity, 20, 236
 Philo of Byzantium, 7
 photoconductor, 273
 photocurrent, 275, 276
 photodetector, 148, 272, 273–78, 280, 289
 photodiode, 273–78
 photoelectric effect, 58, 255–56
 photomultiplier, 272
 photon, 10, 65, 69, 81, 87, 92, 93, 94, 96, 141,
 143, 225–26, 224–28, 259, 256, 261–62, 271,
 272–87
 photon energy, 62
 photon momentum, 67, 70
 photon spin, 141
 Planck, Max, 61, 62, 63, 254, 256
 Planck's constant, 61, 62, 63, 64, 292
 Planck's distribution, 226–28, 259, 269
 Planck's radiation law, 259–62, 264, 266
 Planck's distribution, 260
 polar angle, 118
 population inversion, 284–85
 positron, 135, 141
 potential energy, 13–17, 74, 75, 79, 99, 101, 107,
 117, 136, 144, 163, 227, 246, 247
 principal quantum number, 64, 117, 121, 126, 129
 probability density, 72–73, 76, 78, 91, 95, 105,
 107, 108, 132, 133, 146
 proton
 center of mass behavior, 114–16
 charge, 116
 rest mass, 115, 292
 spin, 141
 proton-electron mass ratio, 292
 Proust, Joseph Louis, 3
 pseudo-momentum. *See* momentum, pseudo
 Ptolemy, 3

Q

quantum confinement, 80–85
 quantum mechanical particle
 ontology, 92–94
 quantum mechanical wave
 ontology, 94

quark, 141
 quartz, 152
 quasi Fermi level, 282, 286, 287
 quasi-particles, 238

R

random walk, 244
 Rayleigh, Lord, 60
 Rayleigh-Jeans model, 60
 recombination, 248, 250
 lifetime, 273
 non-radiative, 273
 radiative, 273, 282
 reduced mass, 115, 116
 refraction, 5
 relativity, 4, 10, 135
 resonance, 19–20
 reverse bias, 248, 249, 272, 275
 reverse leakage, 249, 250
 roots of unity, 161
 Rumford, Count. *See* Thompson, Benjamin
 Rydberg constant, 292
 Rydberg energy, 64, 129, 292

S

saturation current, 249
 Savery, Thomas, 7
 Schrödinger, Erwin, 67
 Schrödinger's wave equation, 66–69, 69–72, 74–
 76, 78, 85, 100, 101, 102, 103, 112, 114, 115,
 119, 125, 126, 129, 135, 146, 158, 162, 167,
 181
 time-dependent, 101
 screening. *See* electron, screening
 second law of thermodynamics. *See*
 thermodynamics, second law of
 semiconductor, 173, 234, 303
 bands, 173, 234–35, 301
 crystal, 149–52
 diode. *See* diode
 doping. *See* doping
 intrinsic, 240
 laser. *See* laser diode
 n-type. *See* doping, n-type
 optoelectronics, 11, 272–87
 p-type. *See* doping, p-type
 separation of variables, 119
 Siemens, Werner von, 60
 simple harmonic oscillator, 21–23, 34, 36, 37, 40,
 227
 Snell's law, 5
 Snellius, Willebrord, 5

solar cell, 1, 148, 152, 209, 272, 273–78
 spdf notation, 125
 specific heat, 8
 speed of light in vacuum, 292
 spherical harmonics, 119–26, 132, 136
 spherical polar coordinates, 118, 260
 spin, 134–43, 171, 221, 298
 electron. *See* electron, spin
 magnetic quantum number. *See* magnetic spin
 quantum number
 neutron. *See* neutron spin
 photon. *See* photon spin
 proton. *See* proton spin
 quantum number, 134
 spin-down, 134, 135
 spin-up, 134
 system, 190–201, 216
 spin excess, 184, 191
 spin-statistics theorem, 135
 spontaneous emission, 62, 266–71, 284
 spring constant, 21
 steam engine, 7, 209
 steam turbine, 7, 209
 Stefan, Josef, 262
 Stefan-Boltzmann law, 262
 stimulated emission, 62, 282–87, 287
 Stirling engine, 209
 Stirling's formula, 186, 187
 stopping voltage, 256

T

temperature, 7, 8, 59, 60, 61, 152, 175, 176, 177,
 178, 195–205, 212, 214, 216, 217, 220, 221,
 223, 224, 225, 227, 232, 245, 259–65, 267–69,
 279, 293, 299, 300, 301, 303, 305, 307
 Celsius, 199
 Fahrenheit, 199
 fundamental, 199, 217
 kelvin, 199
 light bulb filament, 60
 Rankine, 199
 room, 180, 305, 306
 scale, 8, 199
 sun, 60, 269
 tension, 28–29, 34
 Thales of Miletus, 2, 6
 thermodynamics, 1, 7, 9, 11, 182, 199, 200, 216,
 217, 218, 254, 290
 first law of, 8, 209
 second law of, 8, 9, 11, 182, 205–12, 205, 210,
 232, 264, 265
 Thompson, Benjamin, 8
 Thomson, William, 8

transistor, 1, 81, 100, 148, 177
tunneling, 99–112, 113, 144, 145
two-slit experiment, 6, 89–92

U

ultraviolet catastrophe, 61, 261
unit cell, 149, 151, 156, 157, 159, 160, 161, 162,
163, 166, 170, 297
primitive, 151
Wigner-Seitz, 151

V

valence band, 169, 173, 174, 234–35, 235, 238–
39, 245, 248, 249, 273, 279–81, 282, 286, 287,
297, 302, 303, 235–39
valence electron, 139
van Musschenbroek, Pieter, 6
velocity of light, 6, 7, 65, 260, 292
vertical transition, 280
Volta, Alessandro, 6
voltaic cell, 6
von Kleist, Ewald, 6
von Neumann, John, 67, 68

W

Watt, James, 8

wave
monochromatic, 32–36, 69, 86
wave equation
classical. *See* classical wave equation
Helmholtz. *See* Helmholtz wave equation
Schrödinger. *See* Schrödinger's wave equation
wave packet, 236
wave-particle duality, 62, 89, 92, 112
wavevector, 33
Wilkins, John, 7
wurtzite lattice, 150

X

Xu, Wang, 6

Y

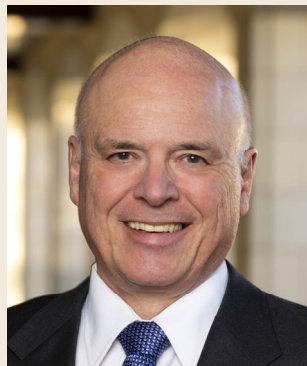
Young, Thomas, 6
Young's slits. *See* two-slit experiment
Yu, Zhu, 6

Z

Zeeman effect, 126
zero-point energy, 80, 81
zinc blende lattice, 149, 150, 151, 171, 235, 297
zinc sulfide. *See* zinc blende lattice

The physical technologies we need for major societal goals, from solar energy to the ways we sense, move, store and process information and design modern materials and devices, are based on ideas and concepts beyond the world of classical physics. A purely classical view gives us no model for the atoms and molecules of chemistry or of material properties such as color or electrical conductivity. We cannot understand the simplest solar cell or even the color and brightness of the sun. To move forward, we need core ideas from quantum mechanics and statistical mechanics, but introducing them often requires multiple courses and textbooks, creating a barrier for a broad range of engineers and scientists to learn these central and powerful ideas. This book introduces all these topics, economically yet rigorously, in one coherent approach suitable for engineers and scientists from a broad range of disciplines. It is well suited both for classroom use and individual study. A large number of quiz questions and problems and solutions are provided. Openly available online lectures are carefully sequenced and linked with the entire main text.

David A. B. Miller (B. Sc., St. Andrews University, Ph.D., Heriot-Watt University) is the W. M. Keck Professor of Electrical Engineering Emeritus and Professor by Courtesy of Applied Physics at Stanford University. Before Stanford, he was with Bell Laboratories from 1981 to 1996. His research has covered optics in digital systems, nanophotonic structures and devices, fundamentals of optics and waves, and semiconductor optics and optoelectronics, including the discovery of the quantum-confined Stark effect and its use for optical modulators throughout modern optical networks.



He has published over 300 scientific papers, holds over 75 patents, has a Google h-index of over 110, and is also the author of the textbook *Quantum Mechanics for Scientists and Engineers* (Cambridge, 2008). In addition to courses on quantum mechanics, optoelectronic devices, and modern physics at Stanford, he has taught open online quantum mechanics classes to over 80,000 students. He has received several awards for his research, is a Fellow of many professional societies, holds two honorary doctorates, and is a Member of the US National Academies of Sciences and of Engineering.

