Modern Physics

Relativity, Quantum Theory, and the Structure of Matter

Martin Veillette

[object Object]

© 2026 Martin Veillette.

This book is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC-BY-NC-SA-4.0). You are free to share and adapt it for non-commercial purposes, with attribution, under the same licence:

https://creativecommons.org/licenses/by-nc-sa/4.0/

This book adapts openly licensed content from several sources. The per-chapter attribution ledger is `SOURCES.md`:

https://github.com/QuadriviumPress/modernPhysics/blob/main/SOURCES.md

The interactive simulations reproduced here as screenshots are published by PhET Interactive Simulations (University of Colorado Boulder) and by the OpenLyceum project under their own licences; each caption links to the running simulation.

Contents

Preface **vi**
0.1 Audience and prerequisites . . . vi
0.2 Organization . . . vi
0.3 How the chapters are built . . . vii
0.4 Editions . . . vii
0.5 License and sources . . . vii

I Relativity **1**

1 The Need for Relativity **2**
1.1 The Ether Hypothesis and Michelson–Morley . . . 3
1.2 Further Null Results: Kennedy–Thorndike and Fizeau . . . 7
1.3 Toward Einstein's Postulates . . . 12
1.4 Summary . . . 16
1.5 Check Your Understanding . . . 17
1.6 Conceptual Questions . . . 18
1.7 Problems . . . 18

2 Special Relativity **21**
2.1 Simultaneity, Time Dilation, and Length Contraction . . . 22
2.2 The Lorentz Transformation and Causality . . . 29
2.3 Doppler Effect, Spacetime Diagrams, and Magnetism . . . 33
2.4 Summary . . . 38
2.5 Check Your Understanding . . . 40
2.6 Conceptual Questions . . . 41
2.7 Problems . . . 41

3 Relativistic Dynamics **45**
3.1 Relativistic Momentum . . . 46
3.2 Relativistic Energy . . . 48
3.3 Four-Vectors and Threshold Energies . . . 49
3.4 Summary . . . 55
3.5 Check Your Understanding . . . 56
3.6 Conceptual Questions . . . 57
3.7 Problems . . . 57

II Wave Optics **60**

4 Interference of Light **61**
4.1 Waves, Huygens's Principle, and Superposition . . . 62
4.2 Young's Double Slit and Coherence . . . 68
4.3 Intensity and Multiple Slits . . . 75
4.4 Thin Films and Interferometers . . . 80
4.5 Summary . . . 89
4.6 Check Your Understanding . . . 90
4.7 Conceptual Questions . . . 90
4.8 Problems . . . 91

5 Diffraction of Light **95**
5.1 Diffraction Basics and the Single Slit . . . 96
5.2 Intensity Patterns and the Double Slit . . . 101
5.3 Diffraction Gratings . . . 105
5.4 Resolution, Bragg Diffraction, and Holography . . . 108
5.5 Summary . . . 115
5.6 Check Your Understanding . . . 116
5.7 Conceptual Questions . . . 116
5.8 Problems . . . 117

III Quantum Theory **121**

6 Particle Properties of Waves **122**
6.1 Blackbody Radiation and Planck's Hypothesis . . . 123
6.2 The Photoelectric Effect and X-Ray Production . . . 126
6.3 Compton Scattering, Pair Production, and the Photon . . . 131
6.4 Summary . . . 136
6.5 Check Your Understanding . . . 138
6.6 Conceptual Questions . . . 138
6.7 Problems . . . 139

7 Wave Properties of Particles **143**
7.1 Matter Waves: de Broglie and Experiment . . . 144
7.2 Wave–Particle Duality and Wave Packets . . . 151
7.3 The Heisenberg Uncertainty Principle . . . 159
7.4 Summary . . . 164
7.5 Check Your Understanding . . . 164
7.6 Conceptual Questions . . . 165
7.7 Problems . . . 165

8 The Schrödinger Equation **169**
8.1 The Wave Function and the Schrödinger Equation . . . 170
8.2 Bound and Scattering States in One Dimension . . . 175
8.3 Tunneling, the Oscillator, and Correspondence . . . 180
8.4 Summary . . . 184
8.5 Check Your Understanding . . . 185

8.6 Conceptual Questions . . . 185
8.7 Problems . . . 186

9 Quantum Mechanics in Three Dimensions 190
9.1 Three Dimensions and Central Potentials . . . 191
9.2 Orbital Angular Momentum . . . 194
9.3 Quantum Numbers and the Isotropic Oscillator . . . 199
9.4 Summary . . . 200
9.5 Check Your Understanding . . . 201
9.6 Conceptual Questions . . . 201
9.7 Problems . . . 202

IV Atoms and Molecules 205

10 The Hydrogen Atom 206
10.1 Energy Levels and Hydrogenic Structure . . . 207
10.2 Wave Functions, Spectra, and Spin . . . 212
10.3 Magnetic Moments and Selection Rules . . . 217
10.4 Summary . . . 222
10.5 Check Your Understanding . . . 222
10.6 Conceptual Questions . . . 223
10.7 Problems . . . 223

11 Many-Electron Atoms 226
11.1 The Exclusion Principle and Screening . . . 226
11.2 Electron Configurations and the Periodic Table . . . 230
11.3 Atomic Spectra, X-Rays, and Lasers . . . 235
11.4 Summary . . . 241
11.5 Check Your Understanding . . . 242
11.6 Conceptual Questions . . . 243
11.7 Problems . . . 243

12 Molecular Structure 246
12.1 Bonding and Hybrid Orbitals . . . 247
12.2 Molecular Orbital Theory . . . 251
12.3 Intermolecular Forces and Molecular Spectra . . . 258
12.4 Summary . . . 264
12.5 Check Your Understanding . . . 265
12.6 Conceptual Questions . . . 265
12.7 Problems . . . 265

V Nuclear and Particle Physics 268

13 Nuclear Physics 269
13.1 Nuclear Structure and Binding Energy . . . 270
13.2 Radioactive Decay . . . 277
13.3 Fission and Fusion . . . 287

13.4 Summary . . . 292
13.5 Check Your Understanding . . . 293
13.6 Conceptual Questions . . . 293
13.7 Problems . . . 294

14 Elementary Particles and the Standard Model 296
14.1 Particles, Forces, and Feynman Diagrams . . . 297
14.2 Accelerators, Leptons, and Quarks . . . 302
14.3 Hadrons, Conservation Laws, and Open Questions . . . 306
14.4 Summary . . . 315
14.5 Check Your Understanding . . . 316
14.6 Conceptual Questions . . . 317
14.7 Problems . . . 317

15 Teaching Guide 320
15.1 Suggested 10-Week Pacing . . . 321
15.2 What To Trim First . . . 321

16 Glossary 322

17 Bibliography 323

Preface

Modern Physics is a first course in the physics that replaced the nineteenth-century picture of space, time, light, and matter. It assumes a year of calculus-based introductory physics — classical mechanics, electromagnetism, waves, and thermodynamics — and uses that foundation to build special relativity, the quantum theory of light and matter, the structure of atoms and molecules, and the physics of nuclei and elementary particles. The book is meant to be readable as a primary text in a one-semester modern-physics course, and usable as a structured reference when those topics appear later.

The material is old enough to be settled and new enough to still feel strange. Special relativity and quantum mechanics are more than a century old; they are also the everyday working language of particle detectors, medical imaging, semiconductor devices, and atomic clocks. The aim here is not to survey every application, but to make the central arguments clear enough that a student who has worked through them can recognize the same physics when it reappears in a research paper, a laboratory, or a later course.

0.1 Audience and prerequisites

The intended reader has completed — or is completing — an introductory calculus-based physics sequence and is comfortable with derivatives, integrals, and elementary differential equations. Linear algebra beyond what usually appears in that sequence is not required; complex numbers appear where quantum mechanics needs them, and are introduced in place. Prior exposure to special relativity or quantum ideas is welcome but not assumed: each part begins where a careful introductory course typically stops.

Instructors will find that the chapters are sized for assignment by numbered section. A typical semester can cover Parts I–III in full and then choose among Parts IV and V according to taste; a more leisurely pace can stop after the hydrogen atom, or press through to the Standard Model. Nothing later in the book silently depends on a chapter that can reasonably be skipped, beyond the dependencies that the text itself flags.

For a compact 10-week pacing guide with suggested section groupings, see Teaching Guide.

0.2 Organization

The book has five parts.

Part I — Relativity begins with the experimental failures of the ether picture, develops special relativity from Einstein's postulates, and closes with relativistic momentum, energy, and the invariant mass that makes particle thresholds computable.

Part II — Wave Optics is a bridge. Interference and diffraction are classical, but they are also the language in which the quantum theory of light and matter is stated. Students who

have already seen this material in an introductory course can move quickly; those who have not will find the chapters self-contained.

Part III — Quantum Theory takes up the particle properties of light, the wave properties of matter, the Schrödinger equation in one dimension, and the extension to three dimensions and angular momentum. This is the core of the course for most instructors.

Part IV — Atoms and Molecules applies that machinery to hydrogen, to many-electron atoms and the periodic table, and to the covalent bond and molecular spectra.

Part V — Nuclear and Particle Physics treats nuclear structure and radioactivity, fission and fusion, and — briefly — the elementary particles and the Standard Model, ending where the open questions begin.

0.3 How the chapters are built

Every chapter opens with learning objectives and a short introduction, then proceeds through numbered reading sections (§*N*.1, §*N*.2, …) kept roughly equal in length so that an assignment can be stated as a range of sections. Worked examples are embedded in the reading, not quarantined at the end. Each chapter closes with a summary and a problem set; some also include conceptual questions. Solutions for every problem are included in the complete edition of the book and omitted from the student edition.

A handful of pedagogical asides appear where they earn their keep: notes for historical context, tips for problem-solving strategy, warnings for common misconceptions, and margin remarks for notation or units that would otherwise interrupt the sentence. Dropdowns hold optional derivations. None of these is a substitute for the main argument; all of them can be ignored on a first reading.

Interactive simulations from PhET and OpenLyceum, locally authored looping animations, and selected demonstration videos are embedded in the web edition wherever motion teaches more than a static picture. In the print and Word editions they become screenshots, generated figures, or poster images, with captions linking to the running source. Every chapter also closes with a static Check Your Understanding review, conceptual questions, and difficulty-tagged problems. A glossary and scholarly bibliography follow the appendices.

0.4 Editions

The website is the primary edition. From the same source the book also builds a complete PDF (with worked solutions), a student PDF (exercises without solutions), a Word document, and a standalone offprint of each chapter. Cross-references, equation numbers, and figure numbers agree across all of them: “Section 4.2” and “Figure 7.3” mean the same thing on the screen, on paper, and in a single-chapter handout.

0.5 License and sources

This book is released under the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International License. It adapts openly licensed material — principally OpenStax *University Physics Volume 3* and *Chemistry 2e*, together with several texts from Benjamin Crowell’s Light and Matter series and Paul D’Alessandris’s *Spiral Modern Physics* — rewritten, reorganized, and extended rather than copied. The per-chapter attribution ledger is the file

`SOURCES.md` in the book's repository. Diagrams and computed figures created for the book share its license; historical photographs and simulations retain the licenses stated in that ledger.

Corrections and suggestions are welcome. An open textbook is a draft that happens to be useful, and it improves when its readers say so.

Part I

Relativity

Chapter 1

The Need for Relativity

Learning Objectives

By the end of this chapter, you should be able to:

- Explain why late-nineteenth-century physics appeared essentially complete, and identify the handful of unresolved puzzles that undid that picture.
- State the principle of relativity as it already existed in Newtonian mechanics, and explain why it does not, by itself, forbid a preferred frame for light.
- State the Galilean transformation and Galilean velocity-addition rule, and explain what they predict for the speed of light.
- Describe the luminiferous-ether hypothesis and the reasoning behind the Michelson–Morley experiment, and compute the fringe shift it predicted.
- Explain why the null result of the Michelson–Morley experiment could not be explained away within Newtonian mechanics and a stationary ether.
- Explain why the Kennedy–Thorndike experiment constrains length-only explanations of the null results and, together with independent clock tests, supports the linked space-and-time effects of Lorentz invariance.
- Describe the Fizeau experiment and explain why its result was equally difficult to accommodate within a simple ether picture.
- State Einstein's two postulates of special relativity and explain why they are jointly incompatible with Galilean relativity.

Introduction

By the 1890s, many physicists believed their subject was nearly finished. Newtonian mechanics predicted the motion of planets, projectiles, and machinery with extraordinary precision; Maxwell's equations had unified electricity, magnetism, and optics into a single theory of the electromagnetic field and correctly predicted that light is an electromagnetic wave. A famous (if perhaps apocryphal) remark attributed to physicists of the era held that all that remained was to measure constants to more decimal places.

That confidence concealed a handful of loose threads. This chapter follows one of them: Maxwell's equations predict a definite speed for light, $c \approx 3.00 \times 10^8$ m/s, but they do not say *relative to what* that speed is measured. Every other wave known to nineteenth-century physics — sound, water waves, waves on a string — is a disturbance in a material medium, and its speed is fixed relative to that medium, not relative to an observer. It was natural to assume light must be the same: a wave in some all-pervading substance, the *luminiferous ether*, with

c being its speed relative to the ether alone. Observers moving through the ether should then measure a different speed for light, just as a swimmer's speed relative to the shore depends on whether they swim with or against a current. Two experiments designed to detect this effect — one that stubbornly refused to find it, and one that found something stranger still — are where this chapter begins, before turning to the radical reinterpretation Einstein proposed for both results at once.

> Today c is *defined* to be exactly 299,792,458 m/s; the meter itself is defined in terms of it. In 1890 it was instead a measured quantity, and Maxwell's equations gave no reason to expect it to be measured the same by every observer.

1.1 The Ether Hypothesis and Michelson–Morley

The Principle of Relativity, Before Einstein

It is worth being precise about what was, and was not, already understood before 1905. Galileo had already argued, nearly three centuries earlier, that uniform motion is undetectable from the inside. In his *Dialogue Concerning the Two Chief World Systems* (1632), he imagined an observer sealed below the deck of a smoothly sailing ship, with no portholes: butterflies fly about the cabin, fish swim in a bowl, drops fall from a bottle into a jar beneath it. Galileo's point was that no mechanical experiment performed in that sealed cabin can tell you whether the ship is at rest in harbor or cruising at constant velocity across a calm sea. The butterflies do not pile up against the stern wall; the drops fall straight down into the jar exactly as they would in harbor. Only *changes* in the ship's velocity — a lurch, a turn — are detectable from inside, because those involve acceleration, and acceleration is not relative in the same way.

This is the **principle of relativity** applied to mechanics alone: the laws of mechanics take the same form in every inertial (non-accelerating) reference frame, so no mechanical experiment, performed entirely inside a closed laboratory, can determine that laboratory's velocity — only changes in velocity are detectable. Formally, this is exactly the content of the Galilean transformation worked out below: Newton's second law, $F = ma$, keeps the same form under $x' = x - vt$ for any constant v, because acceleration is unchanged by the transformation ($a' = a$, since v is constant). Nothing about this principle was controversial in 1890; it was already a working assumption of every mechanic and astronomer.

What was *not* yet settled was whether this principle extended beyond mechanics — in particular, to light and electromagnetism. Maxwell's equations, unlike Newton's laws, single out a specific speed c for electromagnetic waves, and by the reasoning of the Introduction above, a wave's speed is normally measured relative to its medium, not relative to an arbitrary observer. If light truly is a disturbance in a material ether, then Maxwell's equations should hold their simplest form only in the ether's own rest frame, and an observer moving relative to the ether — such as anyone standing on the surface of a planet orbiting the Sun — should be able to detect that motion optically, even though no *mechanical* experiment inside a closed lab could reveal it. The question the rest of this chapter pursues is empirical: can such motion actually be detected?

Galilean Relativity and the Speed of Light

Newtonian mechanics is built on a specific rule for relating measurements made by observers moving relative to one another. Consider two reference frames, S and S', with S' moving at

constant velocity v along the x-axis of S, and with their origins coinciding at $t = 0$. The *Galilean transformation* relates the coordinates of an event as measured in each frame:

$$x' = x - vt, \qquad y' = y, \qquad z' = z, \qquad t' = t. \tag{1.1}$$

The last equation encodes an assumption so natural to everyday experience that it usually goes unstated: time flows identically for every observer, regardless of how they are moving. Differentiating the transformation for x gives the familiar classical velocity-addition rule. If an object moves with velocity u_x in frame S, its velocity in S' is

$$u'_x = u_x - v. \tag{1.2}$$

This is the rule that lets you add your walking speed to a moving walkway, or a thrown ball's speed to the speed of the train from which it is thrown. Applied to light, it makes a sharp prediction: if light travels at speed c relative to the ether, and Earth moves through the ether at orbital speed $v \approx 3 \times 10^4$ m/s, then an observer on Earth measuring light traveling in the same direction as Earth's motion should get $c - v$, and light traveling in the opposite direction should give $c + v$. The effect is small — about one part in 10^4 — but with an interferometer sensitive enough to detect a shift in the interference pattern of order that size, it should be measurable.

> **Warning**
> It is easy to mix up the sign in $u'_x = u_x - v$: the rule subtracts the *frame's* velocity from the object's velocity, not the other way around. A quick check that costs nothing: if S' moves in the $+x$ direction relative to S ($v > 0$) and the object is at rest in S ($u_x = 0$), it must appear to drift in the $-x$ direction in S' ($u'_x = -v$), exactly as the ground appears to slide backward beneath a forward-moving train.

The assumption hiding inside that argument is worth making explicit, because every wave known to nineteenth-century physics encouraged it. A wave is a disturbance of *something*, and that something fixes the frame in which the wave has its textbook speed: Figure 1.1 runs one oscillating source as a ripple on water, as a sound wave in air, and as light. On the first two screens the medium can be put on display directly — the water surface, and the individual air molecules being pushed back and forth — and it is relative to that medium that the disturbance travels at its stated speed. The ether hypothesis is the assumption that the third screen works like the first two, and that the medium there is merely one nobody had yet managed to detect.

The Michelson–Morley Experiment

> **Note**
> Albert Michelson, a graduate of the U.S. Naval Academy, built his first interferometer in 1881 (in Berlin, with support from Alexander Graham Bell) and refined it with Edward Morley at what is now Case Western Reserve University in 1887. Michelson went on to win the 1907 Nobel Prize in Physics — the first American laureate in the sciences — awarded for his precision optical instruments and the spectroscopic and metrological investigations he carried out with them, a body of work that included this interferometer but was not confined to it, even though the experiment's headline result was a *null* one.

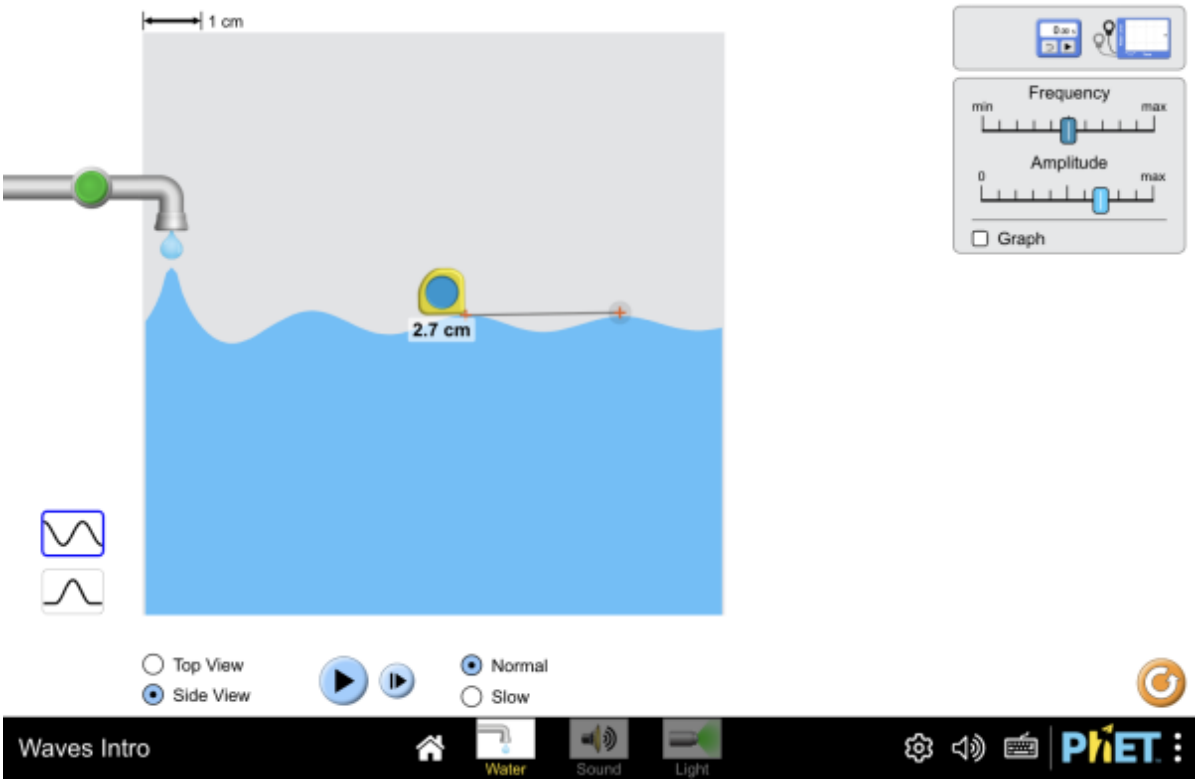

Figure 1.1: A single oscillating source seen as a water wave, a sound wave, and a light wave. For water and sound the medium can be drawn on the screen, and it sets the frame in which the wave travels at its stated speed; light was expected to be no different, and the medium it was supposed to disturb was the ether. Interactive simulation: Waves Intro

Albert Michelson (and later Michelson working with Edward Morley) built exactly such an interferometer. Its principle is to split a beam of light into two perpendicular paths, reflect each off a mirror, recombine the beams, and observe the resulting interference pattern. If Earth moves through the ether, the round-trip travel time along the path parallel to that motion should differ slightly from the round-trip time along the perpendicular path, because a "cross-stream" trip and an "upstream–downstream" trip through a moving medium take different times even when the two paths have equal length — the same effect that makes a boat crossing a river directly and returning take less time than traveling the same distance upstream and back. The predicted difference is small, but the interferometer was sensitive enough to detect a shift of a small fraction of a fringe, and the apparatus was mounted on a stone slab floating on mercury so that it could be rotated smoothly to swap the roles of the two arms.

Computing the Predicted Fringe Shift

The prediction can be made quantitative. Let both interferometer arms have length L, with the apparatus moving at speed v through the ether, and suppose (for the moment) that one arm lies along the direction of motion and the other perpendicular to it. Light traveling down the parallel arm and back takes

$$t_{\parallel} = \frac{L}{c - v} + \frac{L}{c + v} = \frac{2Lc}{c^2 - v^2} = \frac{2L}{c}\left(1 - \frac{v^2}{c^2}\right)^{-1}, \tag{1.3}$$

since it travels at $c - v$ relative to the apparatus going one way (against the ether wind) and $c + v$ coming back. Light traveling down the perpendicular arm and back must be aimed slightly upstream (in the ether frame) to compensate for the sideways drift of the apparatus, so that its actual path through the ether is the hypotenuse of a right triangle with leg L; a short calculation (directly analogous to the light-clock argument of Chapter 2) gives

$$t_{\perp} = \frac{2L}{c}\left(1 - \frac{v^2}{c^2}\right)^{-1/2}. \tag{1.4}$$

> **Tip**
> When a small parameter like v/c appears, expand *before* subtracting, not after. Here $t_{\parallel}$ and $t_{\perp}$ individually differ from $2L/c$ only at order $(v/c)^2$; keeping just one or two terms of the binomial series for each is exactly precise enough to see that they don't cancel completely, while still being simple enough to work with by hand. Trying to compute $t_{\parallel} - t_{\perp}$ exactly first, then approximating, obscures the very cancellation that makes the leading term of order $(v/c)^2$ rather than (v/c).

For $v \ll c$, both expressions can be expanded with the binomial approximation, $(1-x)^{-1} \approx 1 + x$ and $(1 - x)^{-1/2} \approx 1 + x/2$:

$$t_{\parallel} - t_{\perp} \approx \frac{2L}{c}\left[\left(1 + \frac{v^2}{c^2}\right) - \left(1 + \frac{v^2}{2c^2}\right)\right] = \frac{Lv^2}{c^3}. \tag{1.5}$$

Rotating the apparatus by 90° swaps the roles of the two arms, doubling the effect, so the predicted shift in the interference pattern, measured in units of fringes (one fringe corresponding to one extra wavelength λ of path-length difference), is

$$\Delta N = \frac{2(t_{\parallel} - t_{\perp})c}{\lambda} = \frac{2Lv^2}{\lambda c^2}. \tag{1.6}$$

Self-check
Before moving on, check two things in words, not algebra:

- Which arm of the interferometer is longer in the ether picture, and why?
- Why does the first nonzero term in the fringe shift scale like $(v/c)^2$ instead of v/c?

Plugging in the actual 1887 apparatus — effective arm length $L \approx 11$ m (achieved by multiple reflections), sodium light $\lambda \approx 590$ nm, and Earth's orbital speed $v \approx 3.0 \times 10^4$ m/s — gives

$$\Delta N = \frac{2(11\text{ m})(3.0 \times 10^4\text{ m/s})^2}{(590 \times 10^{-9}\text{ m})(3.0 \times 10^8\text{ m/s})^2} \approx 0.37\text{ fringe}, \tag{1.7}$$

comfortably above the apparatus's sensitivity of about 0.01 fringe — roughly a factor of 37 in signal-to-noise. This is the calculation, worked in reverse from the design specifications, that told Michelson and Morley their apparatus was more than sensitive enough for the job.

The Null Result

The prediction was unambiguous: rotating the apparatus should shift the fringe pattern by an amount corresponding to Earth's motion through the ether, roughly 0.4 fringe in the original 1887 apparatus — well above its sensitivity of about 0.01 fringe. The experiment was repeated at different times of day and during different seasons. If a stationary ether existed, Earth's rotation and orbital motion would continually change the direction—and generally also the magnitude—of the laboratory's velocity relative to it. A null result at one moment might therefore reflect an unfavorable orientation or an accidentally small projected velocity, but persistent null results could not. **No shift of the predicted size was ever observed, in any orientation, at any time of year.** The measured shift was consistent with zero and far smaller than the effect Galilean relativity plus a stationary ether required. Later, more sensitive repetitions of the experiment (through the twentieth century, using lasers and eventually rotating optical cavities) have pushed the upper bound on any such effect down by many more orders of magnitude, with the same result: none.

The beam-splitting layout that makes this comparison possible is shown in Figure 1.2.

What a fraction of a fringe looks like, and how little it takes to produce one, can be checked directly in Figure 1.3. Translating one mirror by half a wavelength moves the pattern by one full fringe, so the 0.37 fringe that Michelson and Morley were hunting corresponds to an optical path difference of $0.37\lambda \approx 0.22$ μm, and to a mirror motion of half that — small, but on this display an unmistakable movement of the whole pattern rather than a subtle change of shape. That is why the null result was so hard to argue with.

1.2 Further Null Results: Kennedy–Thorndike and Fizeau

A Second Null Result: The Kennedy–Thorndike Experiment

Michelson–Morley compared equal perpendicular arms while rotating the apparatus. In 1932, Roy Kennedy and Edward Thorndike used unequal arms and searched for changes as Earth's

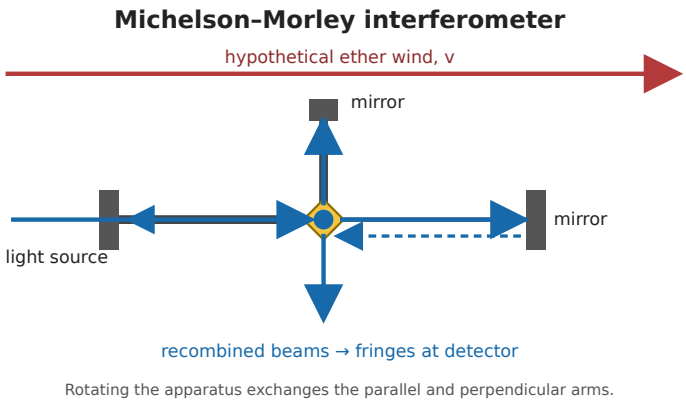

Figure 1.2: The Michelson–Morley arrangement. A beam is split into perpendicular paths and recombined; rotating the apparatus exchanges the arms that are parallel and perpendicular to the hypothesized ether wind. Original schematic by the author.

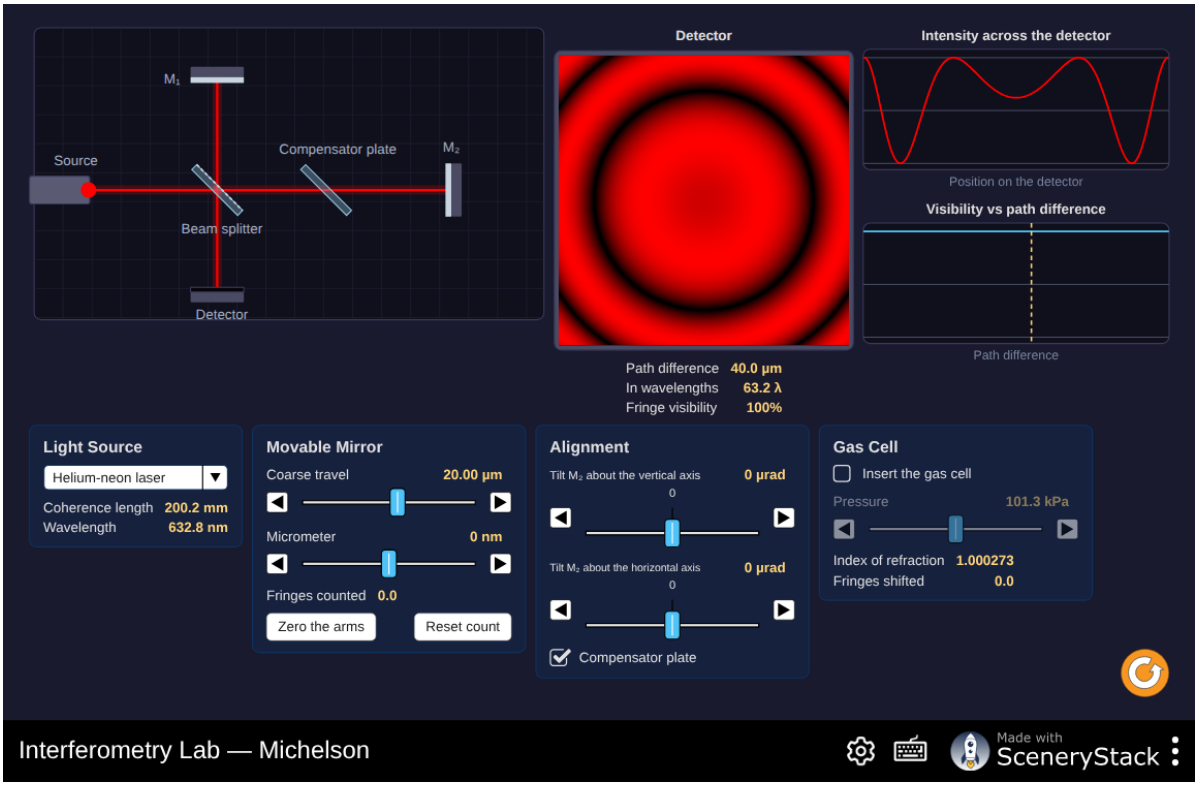

Figure 1.3: The Michelson geometry that Michelson and Morley used, with the arms under direct control. The ether wind is not modeled — no experiment has ever needed it to be — but the instrument's response to a path-length difference between the arms is, and it is that response the 1887 apparatus was calibrated against. Interactive simulation: Interferometry Lab



Figure 1.4: Every measurement of c, including Michelson and Morley's, is necessarily a round-trip measurement — a fact that turns out to matter for exactly the synchronization convention this chapter's null results depend on. Watch on YouTube: Why No One Has Measured The Speed Of Light

velocity relative to a hypothetical preferred frame varied. Their null result closed an important loophole in explanations based on length contraction alone and supported the linked transformation of lengths and clock rates required by Lorentz invariance.

Why unequal arms provide a different test

The surviving photograph in Figure 1.5 shows the scale and mechanical character of the original apparatus described above.

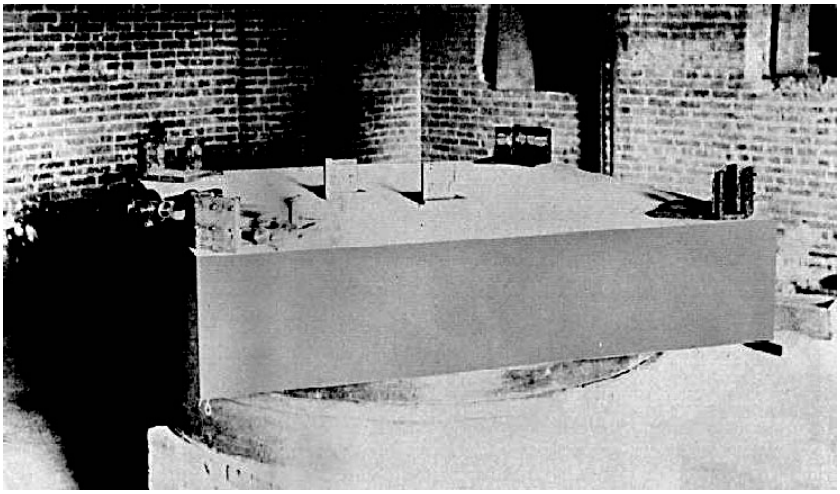

Figure 1.5: The Michelson–Morley apparatus at Case Western Reserve University, circa 1887. Photograph attributed to Case Western Reserve University; public domain via Wikimedia Commons.

One gap in the Michelson–Morley result deserves attention before moving on, because the way it was closed illustrates how tightly interlocking the eventual relativistic explanation would have to be. The Michelson–Morley experiment used two arms of *equal* length, and looked only for a fringe shift as the whole apparatus was *rotated*. This design has a subtle limitation: a length contraction (the Fitzgerald–Lorentz contraction, discussed below in "Attempts to Save the Ether") of exactly the right amount would make $t_{\parallel} = t_{\perp}$ for *any* direction of the apparatus relative to the ether, since whichever arm happens to be aligned with the direction of motion at a given moment contracts by just enough to keep $t_{\parallel}$ equal to $t_{\perp}$ regardless of orientation — so rotating the equal-armed apparatus can never distinguish "no ether effects at all" from "an ether effect perfectly canceled by length contraction alone." It says nothing about whether the *round-trip time itself* depends on velocity through the ether, only about whether the two perpendicular round-trip times remain equal to each other.

Roy Kennedy and Edward Thorndike closed this gap in 1932 with a modified apparatus using two arms of substantially *unequal* length, $L_1 \neq L_2$. With unequal arms, length contraction alone can no longer guarantee $t_1 = t_2$ at all times of year: even if each arm individually contracts as $L\sqrt{1 - v^2/c^2}$, the resulting *difference* $t_1 - t_2$ still depends on the apparatus's instantaneous speed v through the hypothetical ether — a speed that changes over the course of a year as Earth's orbital velocity vector changes direction (and, if the solar system itself moves through the ether, also has a component that does not average to zero). A length-contraction-only patch, with no accompanying change in the rate at which clocks run, predicts a fringe shift that slowly drifts as the year progresses. None was observed, to a precision even finer than Michelson and Morley's original bound.

The Kennedy–Thorndike null result therefore rules out length contraction as a stand-alone patch within this class of ether models. More precisely, it constrains how lengths and the fre-

quency of the light source transform with velocity; interpreted together with Michelson–Morley and independent clock-comparison experiments, it supports the linked length and time effects predicted by the Lorentz transformation. It does not, by itself, constitute a direct measurement of time dilation or uniquely derive the full transformation. This is a useful methodological lesson that recurs throughout physics: a single null result can usually be explained by more than one patch, and it typically takes a *family* of independent experiments, each closing off a different escape route, to establish a genuinely new theory.

A Complication: The Fizeau Experiment

The ether's troubles did not begin with Michelson and Morley. Decades earlier, in 1851, Hippolyte Fizeau performed a related but distinct experiment: rather than looking for an effect of Earth's motion through vacuum, he measured the speed of light *inside a moving medium* — water flowing rapidly through a pipe. Splitting a light beam so that one part traveled with the flow and the other against it, then recombining them interferometrically, Fizeau could measure how much the water's motion changed the light's speed compared to light in stationary water, c/n (where $n \approx 1.33$ is water's refractive index).

> The refractive index n is defined so that light travels at speed c/n in a stationary medium; $n > 1$ for every ordinary transparent material, so $c/n < c$.

The naive, purely mechanical expectation — light as a stream of particles, or the water fully "dragging" the ether along with it, either of which would give ordinary Galilean addition — predicts a shift of exactly v, the full speed of the flowing water:

$$u_{\text{naive}} = \frac{c}{n} + v. \tag{1.8}$$

To first order in the water speed v, Fizeau measured something different, and reproducibly so:

$$u_{\text{measured}} = \frac{c}{n} + v\left(1 - \frac{1}{n^2}\right). \tag{1.9}$$

The water changes the measured speed of light, but only *partially* at this order — by a fraction $f = 1 - 1/n^2$ of its own speed, not the full speed v predicted by simple addition, and not zero either. Augustin-Jean Fresnel had predicted this coefficient in 1818 using an ether model in which a transparent medium partially entrained the ether. The prediction matched Fizeau's measurement, but its physical interpretation required a specialized rule for the interaction between matter and ether.

Warning
Do not read c/n here as if Galilean subtraction had already been applied. That is the light speed in the water's rest frame; the moving-water term is a first-order correction on top of that baseline.

Self-check
If the water is at rest ($v = 0$), what does the measured-speed formula reduce to? What happens to the drag coefficient $1 - 1/n^2$ if $n = 1$?

Fizeau's result sat alongside the Michelson–Morley null result as a second, independent puzzle: whatever the ether was doing, it was not doing anything as simple as being either fully dragged along by matter or entirely undisturbed by it. As will become clear in Chapter 2, once relativistic velocity addition replaces the Galilean rule, Fresnel's fractional drag coefficient emerges automatically — with no separate assumption about entrainment at all — as the leading term of $u_x = (u'_x + v)/(1 + u'_x v/c^2)$ evaluated at $u'_x = c/n$. What looked like an ad hoc patch to the ether model instead emerges, to first order in v/c, from relativistic kinematics.

Extracting the Fresnel coefficient from relativistic velocity addition

This chapter has not yet derived the relativistic velocity-addition formula — that is the work of Chapter 2 — but it is worth seeing, in advance, exactly how the "unmotivated" Fresnel coefficient falls out of it, since the calculation is short. Substitute $u'_x = c/n$ (light's speed in water at rest) into $u_x = (u'_x + v)/(1 + u'_x v/c^2)$:

$$u_x = \frac{\dfrac{c}{n} + v}{1 + \dfrac{v}{nc}}. \tag{1.10}$$

For $v \ll c$, expand the denominator with $(1 + x)^{-1} \approx 1 - x$:

$$u_x \approx \left(\frac{c}{n} + v\right)\left(1 - \frac{v}{nc}\right) \approx \frac{c}{n} + v - \frac{v}{n^2} + O(v^2) = \frac{c}{n} + v\left(1 - \frac{1}{n^2}\right), \tag{1.11}$$

where the cross term $v^2/(nc)$ from the first factor's v times the second factor's $-v/(nc)$ has been dropped as second order in v/c. This reproduces Fizeau's measured formula to first order in v/c, with the fractional drag coefficient $f = 1 - 1/n^2$ appearing automatically — not as a separate hypothesis about how much ether a medium entrains, but as the leading correction to ordinary velocity addition once the relativistic addition rule is used in place of the Galilean one. Higher-precision treatments must also account for effects such as the variation of refractive index with wavelength.

Fizeau is shown in a late-nineteenth-century portrait in Figure 1.6.

The three possibilities are summarized visually in Figure 1.7: no drag, full Galilean drag, and the intermediate result actually observed.

1.3 Toward Einstein's Postulates

Attempts to Save the Ether

The null result did not immediately convince anyone to abandon the ether; instead, it triggered a series of increasingly strained patches:

- **Ether drag.** Perhaps Earth drags the nearby ether along with it, so light near Earth's surface shows no relative motion. The simplest version—complete, uniform drag—conflicts with *stellar aberration*, the small annual shift in the apparent position of stars (about 20.5 arcseconds, first measured by James Bradley in 1727). In the falling-rain analogy, complete local drag would make the incoming light share Earth's sideways motion and would remove the need to tilt the telescope, contrary to observation. Historically, more elaborate drag models were proposed to recover aberration, so aberration alone did not dispose of every possible ether theory; rather, it forced those theories to become substantially more complicated.

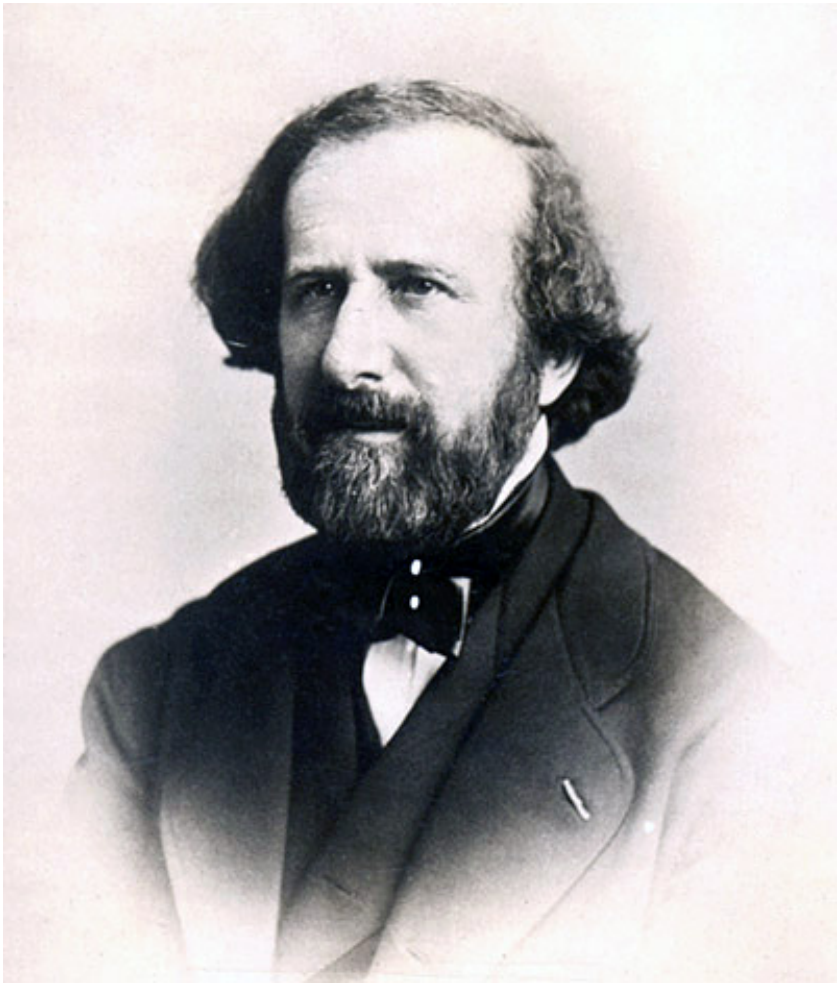

Figure 1.6: Hippolyte Fizeau (1819–1896). Photograph by Charles Reutlinger; Académie des Sciences / Smithsonian Institution Libraries; public domain via Wikimedia Commons.

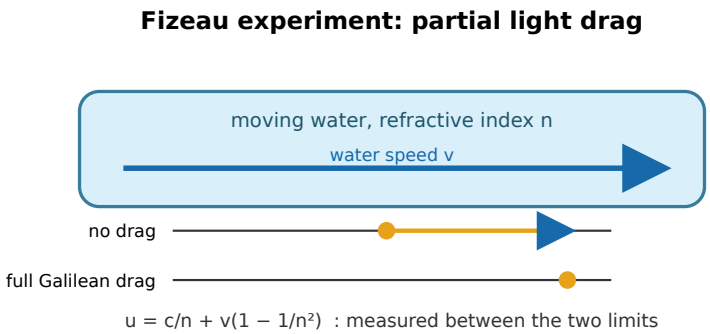

Figure 1.7: Fizeau's moving-water result. The measured velocity is shifted by the Fresnel factor $1 - 1/n^2$, lying between the predictions of no drag and full Galilean addition. Original schematic by the author.

- **The Fitzgerald–Lorentz contraction.** George FitzGerald and, independently, Hendrik Lorentz proposed that objects moving through the ether physically contract along their direction of motion by just the factor needed to make the two arms of the interferometer take equal time, hiding the effect. Concretely, if a rod of ether-frame length L_0 oriented along the direction of motion instead has length $L_0\sqrt{1 - v^2/c^2}$, then $t_\parallel$ recomputed with this shortened arm length equals $t_\perp$ exactly, and no fringe shift results, for any v. Lorentz went further, introducing a fictitious "local time" that varied with position in exactly the combination $t - vx/c^2$ that would, decades later, turn out to be the correct relativistic time coordinate. Taken on its own, however, this is an *ad hoc* fix — a contraction (and a local time) invented for the sole purpose of explaining away one null result, with the ether itself still doing the conceptual work of defining a single true rest frame. It turns out to contain a genuine piece of the truth, but only once it is derived from a deeper principle rather than posited to save a failing hypothesis, which is what Einstein did in 1905 (Chapter 2).
- **Emission theories.** Perhaps light simply travels at speed c relative to its *source*, with no ether needed at all — a proposal in the spirit of a bullet's speed depending on the gun that fired it. De Sitter argued in 1913 that a simple version of this theory would badly distort observations of binary stars: light emitted while a star approached Earth would arrive faster than light emitted while it receded, potentially scrambling the apparent orbit. Such distortions are not observed, providing important evidence against simple ballistic theories. The historical argument was not conclusive by itself, because interactions with intervening matter might erase a source-dependent speed before the light reached Earth; laboratory and astronomical tests taken together provide the stronger case that light's vacuum speed is independent of its source.

Worked Example: Estimating the Aberration Angle

The stellar-aberration bullet above asserts a value of about 20.5 arcseconds; it is worth seeing where a number of that size comes from, since the same reasoning about combining velocities recurs throughout this chapter. Consider a star directly "overhead," so that in the Sun's rest frame its light arrives traveling straight down (speed c, along $-y$) at a telescope. In the frame of Earth, which moves at orbital speed $v \approx 3.0 \times 10^4$ m/s in some direction in the x-y plane perpendicular to the line of sight, ordinary (Galilean) velocity subtraction gives the light's velocity components as measured on Earth: still $-c$ along y, but now also $-v$ along x, exactly as raindrops falling straight down acquire an apparent forward velocity component for a runner moving beneath them. The apparent direction of arrival is tilted from vertical by an angle θ with

$$\tan\theta = \frac{v}{c} = \frac{3.0 \times 10^4 \text{ m/s}}{3.0 \times 10^8 \text{ m/s}} = 1.0 \times 10^{-4}. \tag{1.12}$$

> The ratio $v/c \sim 10^{-4}$ for Earth's orbital motion is the recurring small number of this chapter: it sets the size of the Michelson–Morley fringe shift, the aberration angle, and the fractional Galilean speed shift in Exercise 1 alike.

Since this is a very small angle, $\theta \approx 1.0 \times 10^{-4}$ rad. Converting to arcseconds (using 1 rad = 206,265″),

$$\theta \approx (1.0 \times 10^{-4}\ \text{rad}) \left(\frac{206{,}265''}{1\ \text{rad}} \right) \approx 20.6'', \tag{1.13}$$

in close agreement with the historically measured value of $20.5''$. Because Earth's velocity vector sweeps around in a full circle over the course of a year (always perpendicular to the instantaneous direction to a star near the pole of the ecliptic), each such star traces out a small circle of this same angular radius on the sky annually — exactly the periodic wobble Bradley detected in 1727, long before anyone suspected the ether was in trouble.

Each patch could account for the existing data but only by adding a new, unmotivated assumption, and no single patch accounted for *both* puzzles — the Michelson–Morley null result and the Fizeau partial-drag coefficient — without contradicting itself. What was needed was not another patch but a new starting point.

Einstein's Postulates

In 1905, Einstein proposed to stop patching Galilean relativity and instead take the null result at face value, elevating it (together with the principle of relativity already implicit in mechanics, discussed above) to a postulate [Einstein, 1905b]. Special relativity rests on two statements:

1. **The principle of relativity.** The laws of physics take the same form in all inertial reference frames. No experiment performed entirely within a closed laboratory can distinguish one inertial frame from another.

2. **The constancy of the speed of light.** The speed of light in vacuum has the same value c in every inertial reference frame, independent of the motion of the source or the observer.

The first postulate was not new — as Galileo's ship illustrates, it already held for Newtonian mechanics, where no mechanical experiment can detect uniform motion (this is why you cannot feel the constant-velocity phase of a smooth flight, and why the butterflies in Galileo's cabin fly about undisturbed). Einstein's step was to insist that this postulate apply to *all* of physics, electromagnetism included, so that no experiment of any kind — mechanical or optical — can distinguish inertial frames; there is no preferred, ether-defined rest frame for anything, including light. The second postulate is the one that clashes head-on with Galilean relativity: if $u'_x = u_x - v$ held for light as it does for baseballs, two observers in relative motion could not both measure the same speed c for the same light beam. Taken together, the postulates require that time and space themselves — not just the "aether wind" — behave differently than Galilean relativity assumes.

> **Self-check**
> If two observers are moving relative to each other, can they both keep the Galilean velocity-addition rule *and* measure the same vacuum speed of light? If not, which assumption has to change?

This is a genuinely radical move. Rather than asking "what medium is light waving in?", Einstein asked what kinematics — what set of rules for relating space and time coordinates between observers — is consistent with light having the same speed for everyone. The answer

is not the Galilean transformation but the *Lorentz transformation*, and working out its consequences — the relativity of simultaneity, time dilation, length contraction, and (as noted above) the Fizeau drag coefficient as a special case of velocity addition — is the subject of Chapter 2.

> **See Also**
> Special Relativity derives the Lorentz transformation from these two postulates and works out its consequences in detail — time dilation and length contraction (the effects this chapter's ether patches had to invent piecemeal), the relativity of simultaneity, and the relativistic velocity-addition formula used above to recover the Fizeau coefficient.

Einstein was not working in a vacuum: Lorentz had already written down transformation equations of essentially the same mathematical form (motivated, as noted above, by fitting the Michelson–Morley and Fizeau data rather than by a physical principle), and Henri Poincaré, in the same year, independently emphasized that the principle of relativity should be regarded as an exact law of nature rather than an approximate consequence of ether dynamics, and showed that the Lorentz transformations form a mathematical group — a property essential for consistency, since applying two velocity boosts in succession must itself be equivalent to some single valid transformation. What distinguished Einstein's paper was not the equations themselves but their derivation: rather than reverse-engineering a transformation to fit the data, he started from the two postulates and showed that the Lorentz transformation is the *unique* kinematic consequence of taking them seriously, with time dilation, length contraction, and the relativity of simultaneity following as necessary theorems rather than independently adjustable assumptions.

It is worth appreciating just how much this single change in starting point accomplishes. The Michelson–Morley null result, stellar aberration, binary-star observations, and the Fizeau drag coefficient were separate empirical constraints that simple ether and emission models struggled to accommodate together. Relativistic kinematics provides a common framework for all four, without assigning light to a material medium or introducing a separate ether parameter for each result.

1.4 Summary

- The principle of relativity — that no mechanical experiment inside a closed laboratory can detect uniform motion — already applied to Newtonian mechanics before 1905 (Galileo's ship). What was not settled was whether it extended to optics and electromagnetism, where Maxwell's equations single out a specific speed c naturally interpreted as a speed relative to a medium, the luminiferous ether.
- The Galilean transformation and its associated velocity-addition rule, $u'_x = u_x - v$, predict that an observer moving through the ether should measure a speed of light different from c, by an amount of order Earth's orbital speed; the predicted Michelson–Morley fringe shift is $\Delta N = 2Lv^2/(\lambda c^2) \approx 0.4$ fringe for the 1887 apparatus.
- The Michelson–Morley interferometer was sensitive enough to detect this predicted shift but found none, at any time of year or orientation: the speed of light showed no dependence on the observer's motion through the presumed ether.
- The Kennedy–Thorndike experiment, using unequal interferometer arms, constrained length-contraction-only explanations. Together with Michelson–Morley and independent

clock tests, its null result supports the linked length and time effects of Lorentz invariance.
- The Fizeau experiment, measuring the speed of light in flowing water, found to first order a *partial* drag coefficient, $f = 1 - 1/n^2$ — neither the full entrainment nor the zero entrainment that simple ether pictures would predict. Relativistic velocity addition produces this coefficient naturally at first order in v/c.
- Proposed rescues of the ether hypothesis faced additional constraints: stellar aberration conflicts with simple complete-drag models, binary-star observations weigh against simple emission theories, and length contraction introduced solely to hide the Michelson–Morley effect was an incomplete patch. More elaborate models could evade individual objections, but only at the cost of additional assumptions.
- Einstein's 1905 postulates take the null result at face value: the laws of physics (postulate 1) and the speed of light in particular (postulate 2) are the same in every inertial frame. Together these are incompatible with the Galilean transformation and require a new kinematics, developed in Chapter 2, which reproduces all of the results above — Michelson–Morley, aberration, binary-star timing, and the Fizeau coefficient — without any adjustable ether parameters.

1.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens with the Michelson–Morley apparatus.

1. **Multiple choice, from a figure.** Why are the Michelson–Morley interferometer's two perpendicular arms essential to the ether-wind test? Figure description: Schematic of the Michelson–Morley interferometer with perpendicular light paths, a beam splitter, mirrors, and a detector. Choices: (A) Rotating the apparatus swaps the arms' orientation relative to the proposed ether wind, so a directional travel-time difference would change. (B) They force light to travel faster in one arm than in the other. (C) They make one beam pass through water while the other remains in air. (D) They remove the need to compare the phases of the returning beams.

2. **True or false.** To first order in the water speed divided by c, relativistic velocity addition reproduces Fizeau's partial-drag coefficient 1 - $1/n^2$.

3. **Drag the words.** Complete the contrast between Galilean kinematics and the optical null result. The Galilean rule gives $u'_x =$ _____. It therefore predicts an _____ for a moving laboratory, whereas Michelson–Morley found _____. Einstein instead required every inertial observer to measure the same vacuum light speed _____. (Terms: $u_x - v$; ether wind; no systematic fringe shift; c; $u_x + v$; complete ether drag.)

4. **Fill in the blanks.** Complete the comparison between the predicted and observed Michelson–Morley signals. For the 1887 apparatus, a stationary-ether model predicted about _____ fringe, while rotation produced _____ systematic fringe shift.

5. **Mark the words.** Mark the two statements that form Einstein's postulates. "The laws of physics have the same form in every inertial frame, and the vacuum speed of light is the same for every inertial observer. An ether wind and absolute uniform motion are therefore not observable."

1.6 Conceptual Questions

1. A sealed laboratory moves uniformly through space. Explain why no internal mechanical experiment can determine the laboratory's absolute speed.

2. Why would a stationary luminiferous ether single out one inertial frame, in conflict with the principle of relativity?

3. The Michelson interferometer compared two perpendicular light paths. Why was rotating the entire apparatus essential to the proposed ether-wind test?

4. Fizeau measured light in moving water. Why did partial dragging challenge both complete ether drag and a completely stationary ether?

5. Which assumption must fail if both Galilean velocity addition and invariant light speed cannot hold at once?

1.7 Problems

Exercise 1.1
(Straightforward)

Earth orbits the Sun at approximately $v = 3.0 \times 10^4$ m/s. Using the (incorrect) Galilean velocity-addition rule, estimate the fractional difference $(c + v)/(c - v) - 1$ between the speed of light measured "downstream" and "upstream" of Earth's motion through a stationary ether. Compare your estimate to the sensitivity you would need in an experiment to detect it.

Exercise 1.2
(Straightforward)

A swimmer who swims at speed u in still water crosses a river of width L flowing at speed $v < u$. (a) Find the time to swim straight across and back if the swimmer aims to always move perpendicular to the bank as seen from the shore. (b) Find the time to swim a distance L directly upstream and then back downstream. (c) Show the two times are unequal, and explain the analogy to the two arms of the Michelson–Morley interferometer.

Exercise 1.3
(Straightforward)

Explain, in your own words, why a null result from a single run of the Michelson-Morley experiment would not have been convincing on its own, and why repeating it at different times of year strengthened the conclusion.

Exercise 1.4
(Straightforward)

Suppose light, rather than obeying Einstein's second postulate, were emitted at speed c relative to its source (an "emission theory" of light), like a bullet fired from a moving gun. Explain qualitatively why light from the two stars of a binary system, observed from Earth over one orbital period, would arrive with systematically distorted timing under this theory, and why this distortion is not observed.

Exercise 1.5
(Moderate)

Explain why the principle of relativity (postulate 1) already applied to Newtonian mechanics before 1905, and identify precisely what Einstein's second postulate adds that Newtonian mechanics did not have.

Exercise 1.6
(Moderate)

Derive the fringe-shift formula $\Delta N = 2Lv^2/(\lambda c^2)$ by working through the steps in the text, and use it to estimate the fringe shift a modern Michelson–Morley-style apparatus with arm length $L = 4$ km (comparable to a gravitational-wave observatory) and $\lambda = 1064$ nm would predict for Earth's orbital speed, if a stationary ether existed. Comment on how much more sensitive such an apparatus is, in principle, than Michelson and Morley's original tabletop instrument.

Exercise 1.7
(Moderate)

A modern, high-precision version of the Michelson–Morley experiment finds an upper bound on any fringe shift roughly 10^{-5} times smaller than the shift 0.4 fringe originally predicted by Galilean relativity plus a stationary ether. If this bound were instead interpreted (incorrectly, but for the sake of the exercise) as an upper limit on Earth's residual speed through a stationary ether, using the same fringe-shift formula, what upper limit on that speed would it imply? Compare this to Earth's orbital speed.

Exercise 1.8
(Moderate)

Using the Fizeau formula $u_{\text{measured}} = c/n + v(1 - 1/n^2)$, compute the predicted speed (in the lab frame) of light traveling through water ($n = 1.33$) flowing at $v = 7.0$ m/s, both with and against the light's direction of travel. Find the difference between these two speeds, and compare it to the difference $2v$ that simple Galilean addition (full dragging) would predict.

Exercise 1.9
(Moderate)

Jupiter orbits the Sun at an orbital speed of about $v = 1.3 \times 10^4$ m/s. If an astronomer stationed on Jupiter observed the same "overhead" star discussed in the worked example on aberration, using the same reasoning, what angular radius (in arcseconds) would that astronomer measure for the star's annual aberration ellipse? Explain why the ratio of your answer to Earth's $20.5''$ depends only on the ratio of the two planets' orbital speeds.

Exercise 1.10
(Challenging)

Explain why an ether model in which Earth *fully* drags the surrounding ether along with it (proposed to explain the Michelson–Morley null result) is inconsistent with the existence of stellar aberration, using the falling-rain analogy described in the text: a person standing still under vertically falling rain gets wet only from directly above, while a person running through the rain must tilt an umbrella forward to stay dry. What would "full ether drag" have to imply about the umbrella, if starlight is the falling rain and Earth's atmosphere is analogous to the region being dragged?

Exercise 1.11
(Challenging)

Explain, in your own words, why the original Michelson–Morley design (equal-length arms, apparatus rotated) cannot by itself distinguish "no ether effects" from "an ether effect exactly canceled by a velocity-dependent length contraction." Then explain why making the two arms unequal in length (Kennedy–Thorndike) closes this loophole.

Exercise 1.12
(Challenging)

Suppose (contrary to fact) that the Kennedy–Thorndike experiment *had* detected a fringe drift over the course of a year, of a size consistent with length contraction being the *only* relativistic effect (i.e., with clocks ticking at the same rate in every frame). Explain qualitatively why this outcome would have been logically consistent with the original Michelson–Morley null result, even though it would have been inconsistent with the time dilation and Lorentz transformation developed in Chapter 2.

Chapter 2

Special Relativity

Learning Objectives

By the end of this chapter, you should be able to:

- Derive and apply the relativity of simultaneity from Einstein's postulates, and describe an operational procedure for synchronizing clocks within a single frame.
- Derive and apply time dilation, including the concept of proper time.
- Derive and apply length contraction, including the concept of proper length.
- Use the Lorentz transformation to relate the space and time coordinates of an event in two inertial frames.
- Apply the relativistic velocity-addition formula, show that it never yields a result exceeding c, and use it to explain the Fizeau drag coefficient of Chapter 1.
- Construct and interpret simple spacetime diagrams, including the light cone and its role in cause and effect.
- Explain the resolution of the twin paradox in terms of the asymmetry between inertial and non-inertial motion.
- Derive and apply the relativistic Doppler effect.
- Explain why magnetism can be understood as a relativistic consequence of length contraction acting on electric fields.

Introduction

Chapter 1 established that Einstein's two postulates — the equivalence of all inertial frames, and the invariance of the speed of light — are incompatible with the Galilean transformation. This chapter works out what *does* follow from the postulates: a new set of rules relating space and time measurements between observers in relative motion, called the Lorentz transformation. Its consequences are strange by the standards of everyday experience — moving clocks run slow, moving objects are measured as shortened, and two events that are simultaneous for one observer need not be simultaneous for another — but they are not arbitrary. Each follows directly, and only, from insisting that every inertial observer measures the same speed c for light.

Throughout, an **event** is something that happens at a definite place and a definite time — a flashbulb going off, a particle passing a marker — specified by four coordinates (x, y, z, t) in a given reference frame. Two different inertial observers, in relative motion, will in general assign different coordinates to the same event; the question this chapter answers is exactly how those coordinate sets are related.

2.1 Simultaneity, Time Dilation, and Length Contraction

The Relativity of Simultaneity

Consider a train car moving at constant velocity v relative to the ground, with a light source at its exact center. When the source flashes, light travels outward in all directions at speed c — in *every* inertial frame, by the second postulate.

To an observer sitting at the center of the car, the light reaches the front and back walls simultaneously, since both walls are equidistant from the source and light travels at the same speed c in both directions in the car's own frame.

To an observer standing on the ground watching the car go by, the situation is different. In the ground frame, light still travels at speed c in both directions — but the back wall of the car is moving *toward* the point where the backward-going light was emitted, while the front wall is moving *away* from the point where the forward-going light was emitted. So light reaches the back wall first. The two arrivals, simultaneous in the car's frame, are **not** simultaneous in the ground frame.

This is not a measurement error or a signal-delay artifact to be corrected for — it is a genuine disagreement about which events are simultaneous, forced on us by the requirement that both observers measure the same speed c for the same light pulses. **Simultaneity is relative to the observer's frame of motion**, not an absolute, frame-independent relation between events. This single fact is the seed from which time dilation and length contraction both grow.

The basic geometry is shown in Figure 2.1. The two flashes are simultaneous in the train frame, but the moving observer travels toward the front flash and away from the rear flash.

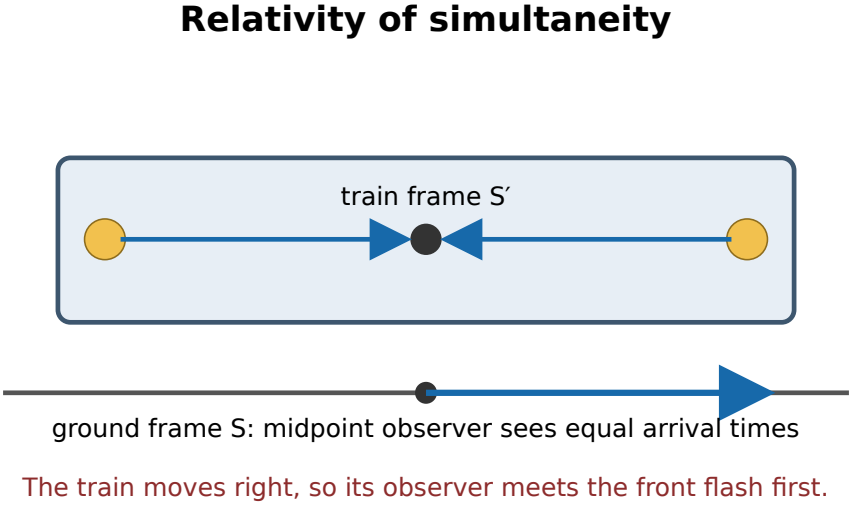

Figure 2.1: Simultaneity depends on the observer's frame. The diagram is schematic: the ground observer is at the midpoint, while the train observer moves toward the front flash. Original schematic by the author.

An Operational Procedure for Synchronizing Clocks

The train example shows *that* simultaneity depends on the observer's frame, but it is worth being explicit about *how* an observer within a single frame decides that two clocks, at different locations, are synchronized in the first place — otherwise "simultaneous in the car's frame" is just as vague a phrase as "simultaneous" was before Einstein.

Here is one way to do it, entirely within a single inertial frame S. An observer, Amy, sits at a fixed location with a clock. At some time t_1 (her own clock reading), she sends a light pulse toward a distant mirror a known distance away. The pulse reflects and returns to her at time t_3. Because light travels at the same speed c in both directions (postulate 2), the pulse must have reached the mirror at the midpoint time,

$$t_2 = \frac{t_1 + t_3}{2}. \tag{2.1}$$

Amy can use this fact to synchronize a second clock, placed at the mirror's location: she instructs whoever is stationed there to set their clock to read t_2 at the moment the light pulse arrives. Any two clocks synchronized this way — by "radar," bouncing a light signal and splitting the round-trip time evenly — will agree with each other and with any other pair of clocks in the same frame synchronized by the same method. (An equivalent, more homely method: synchronize two clocks side by side, then separate them by moving them apart *very slowly*; since relativistic effects on a moving clock's rate are of order v^2/c^2, they can be made as small as desired by transporting the clocks slowly enough.)

This procedure defines, unambiguously, a notion of "simultaneous" *within* a single inertial frame — but it is frame-dependent by construction, because it relies on the frame in which the observer doing the synchronizing is at rest. An observer in a second frame S', moving relative to S, will synchronize clocks by the identical procedure applied within S', and — as the train example already showed — the two resulting sets of "simultaneous" clocks will not agree with each other. Both sets of clocks are perfectly well synchronized, each within its own frame; there is simply no frame-independent fact about which distant events are simultaneous with a given event here and now.

Time Dilation

Consider a **light clock**: two mirrors facing each other, separated by a distance L, with a single light pulse bouncing between them. In the frame where the clock is at rest, one "tick" — one round trip — takes

$$\Delta t_0 = \frac{2L}{c}. \tag{2.2}$$

This time, measured in the frame where the clock does not move, is called the **proper time** between the two events (pulse leaves the bottom mirror; pulse returns to the bottom mirror) — the time measured by a clock present at both events, at the same location.

Now view the same clock from a frame in which it moves at speed v perpendicular to the line joining the mirrors. In this frame, between emission and return the light pulse must travel not just up and down but also sideways with the clock, tracing a diagonal path. Since light still travels at speed c in this frame (postulate 2), and the diagonal path is longer than $2L$, the round trip must take *longer* as measured here. Writing Δt for the round-trip time in this frame, the pulse travels a horizontal distance $v\Delta t$ while covering the diagonal distance $c\Delta t$, and by the Pythagorean theorem applied to each mirror-to-mirror leg,

$$\left(\frac{c\Delta t}{2}\right)^2 = L^2 + \left(\frac{v\Delta t}{2}\right)^2. \tag{2.3}$$

Substituting $L = c\Delta t_0/2$ and solving for Δt gives

$$\Delta t = \frac{\Delta t_0}{\sqrt{1 - v^2/c^2}} \equiv \gamma\, \Delta t_0, \tag{2.4}$$

where the **Lorentz factor**

$$\gamma \equiv \frac{1}{\sqrt{1 - v^2/c^2}} \tag{2.5}$$

is greater than or equal to 1 for any $v < c$, and grows without bound as $v \to c$. This is **time dilation**: a clock moving at speed v relative to an observer is measured by that observer to run slow, ticking out $\Delta t = \gamma \Delta t_0$ of the observer's own time for every Δt_0 of proper time it displays. The effect is symmetric — each of two observers in relative motion sees the *other's* clock as running slow, since each is equally entitled to regard themselves as at rest. Although derived here for a light clock, time dilation applies to time itself, and hence to every physical process — mechanical clocks, radioactive decay rates, biological aging — since two different physical clocks, built differently, must stay in agreement in every frame or their disagreement could be used to detect absolute motion, contradicting postulate 1.

> The dimensionless ratio $\beta \equiv v/c$ is standard shorthand throughout relativity; the Lorentz factor is then simply $\gamma = 1/\sqrt{1 - \beta^2}$, and $\beta \to 1$ is the same limit as $v \to c$.

The two clocks of that argument run side by side in Figure 2.2: one at rest in the laboratory, one gliding past at a speed you set, with the moving pulse's zigzag path drawn in. Nothing in the simulation makes the moving clock tick slowly by fiat. Both pulses travel at c; the moving one simply has a longer path to cover between reflections, and the two tick counts drift apart by exactly the factor γ computed above.

Time dilation is not a hypothesis awaiting confirmation; it is routinely observed. Muons created by cosmic rays in the upper atmosphere have a mean lifetime, at rest, of about 2.2 μs — long enough, at nearly the speed of light, to travel only a few hundred meters before decaying, far short of the several kilometers to Earth's surface. Yet large numbers of these muons are detected at sea level. In Earth's frame, the muons' internal "clock" — the process governing decay — runs slow by the factor γ, extending their mean range by that same factor, which is exactly what is observed. (This effect was confirmed with precision in a classic 1941 experiment by Bruno Rossi and David Hall, comparing muon flux measured at the top of Mount Washington in New Hampshire to the flux at a lower elevation.)

The Twin Paradox

Time dilation invites an apparent paradox that is worth confronting directly. Suppose Alice stays on Earth while her twin Bob boards a spaceship, accelerates to a large fraction of c, travels to a distant star, turns around, and returns to Earth. From Alice's perspective, Bob's clock runs slow throughout the trip (by the factor γ), so Bob should be younger than Alice when he returns. But motion is relative — can't Bob equally well claim that *he* was at rest the whole time, and that it was Alice, receding and then approaching, whose clock ran slow? If

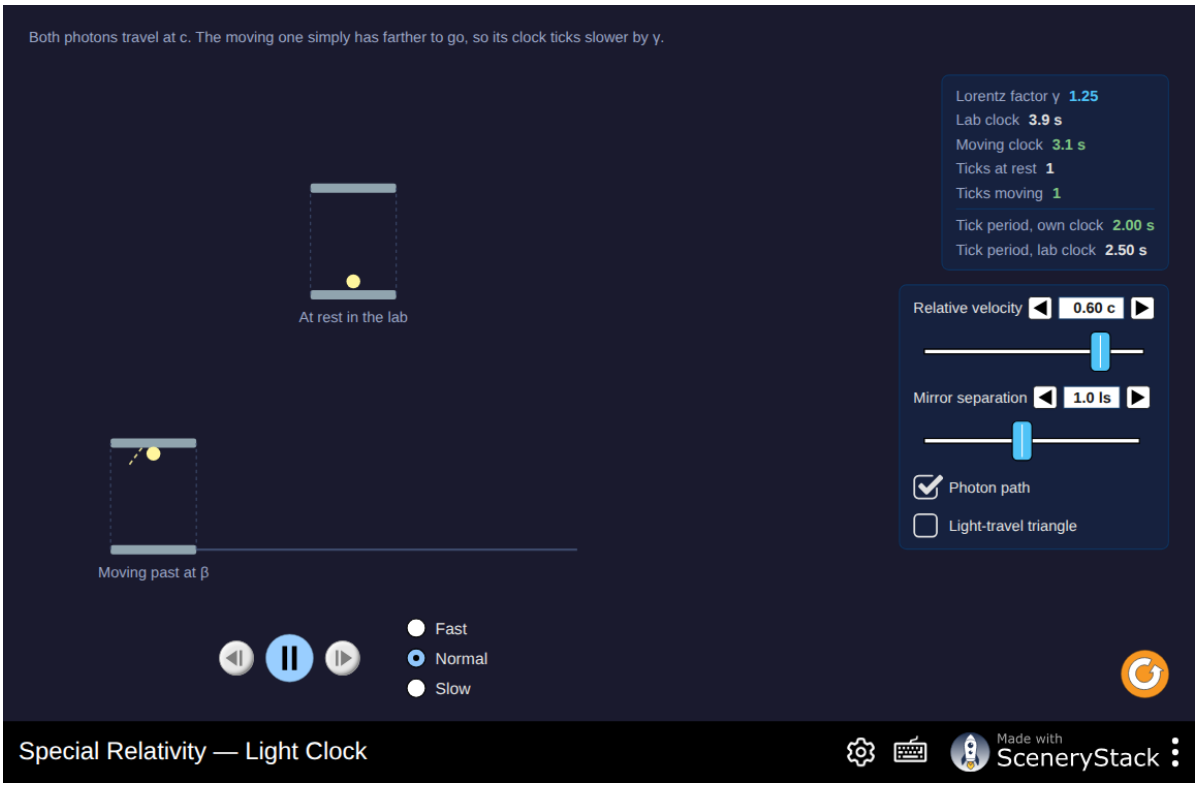

Figure 2.2: A light clock at rest beside an identical one moving at $\beta = v/c$, with the diagonal light path and both tick counts displayed. The other screens of this simulation return in the sections that follow — an interactive Minkowski diagram, the ladder-and-barn paradox (an application of length contraction), the twin paradox, and the relativistic Doppler effect, each one screen further along. Interactive simulation: Special Relativity

so, each twin should conclude the other has aged less, which is a genuine contradiction: when they are reunited at the same place, they can directly compare clocks, and only one answer can be correct.

The resolution is that the situation is **not** symmetric between the twins, and the asymmetry is exactly the thing time dilation was derived for: proper time is measured by a clock that is present at both endpoint events, without changing inertial frames in between. Alice never leaves a single inertial frame. Bob does not: he must decelerate, reverse course, and re-accelerate at the turnaround point (and again to land back on Earth), and no inertial frame is present at all three key moments of his trip — departure, turnaround, and return — the way Alice's single frame is present at both departure and return. Time dilation, as derived above, compares proper time in one frame to elapsed time in *a single other* inertial frame; it does not directly apply to a clock that switches frames partway through. When the trip is analyzed correctly — by computing the proper time elapsed along each twin's actual worldline in a single, fixed inertial frame (say, Earth's) — Bob's path necessarily accumulates less proper time than Alice's straight, inertial path between the same two events, essentially because his path involves a nonzero, non-inertial detour that always makes his elapsed proper time work out to be shorter, not the same. Bob really is younger when he returns, and both twins, correctly accounting for Bob's acceleration, agree on this.

The physical content of the paradox's resolution is this: there is no contradiction, because *only Bob* can locally detect (with an accelerometer, for instance) that he underwent a change of inertial frame. That detectable asymmetry is exactly what breaks the naive symmetry argument and picks out which twin's elapsed time is shorter.

Length Contraction

Time dilation has a companion effect for lengths. Consider a rod of length L_0 at rest along the x-axis of frame S; call L_0 the **proper length** — the length measured in the frame where the rod is at rest. How long is this rod as measured by an observer in frame S', moving at speed v relative to S along the rod's length?

> **Tip**
> Before substituting into $\Delta t = \gamma \Delta t_0$ or $L = L_0/\gamma$, identify explicitly which observer measures the *proper* quantity: Δt_0 belongs to whichever frame has both events happen at the same location, and L_0 belongs to whichever frame the object is at rest in. Every other frame's value is obtained from that one by multiplying or dividing by γ — never the reverse — and mixing up which frame is "proper" is the most common source of a factor-of-γ error in these problems.

To measure a moving rod's length, an observer must record the positions of both ends *at the same instant* in their own frame. Consider the muon example again, now from the muon's own rest frame, in which Earth's atmosphere rushes past at speed v. During the muon's mean proper lifetime Δt_0, a layer of atmosphere of thickness $L = v\Delta t_0$ passes it. In Earth's frame, that same layer is at rest, so its proper thickness is $L_0 = v\Delta t$, where $\Delta t = \gamma \Delta t_0$ is the muon's mean lifetime in that frame. Therefore,

$$L = v\Delta t_0 = \frac{v\Delta t}{\gamma} = \frac{L_0}{\gamma}, \tag{2.6}$$

or

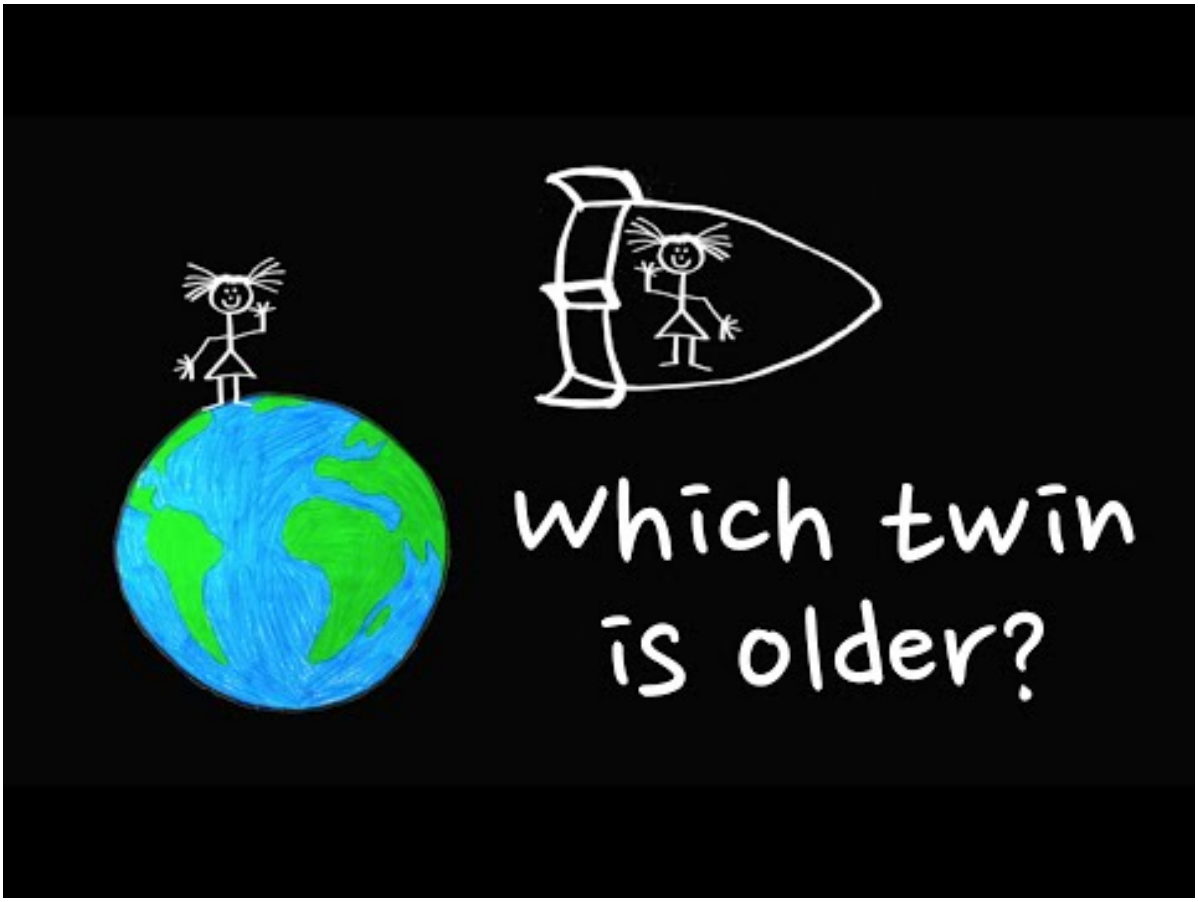

Figure 2.3: A visual companion to the argument above, tracking each twin's worldline through the turnaround to show exactly where the apparent symmetry breaks. Watch on YouTube: Special Relativity and the Twin Paradox

$$L = \frac{L_0}{\gamma} = L_0\sqrt{1 - v^2/c^2}. \tag{2.7}$$

This is **length contraction**: an object of proper length L_0, measured by an observer relative to whom it moves at speed v along its own length, is found to have length $L = L_0/\gamma \leq L_0$. Only lengths *along* the direction of relative motion contract; lengths perpendicular to the motion are unaffected (a consequence one can show is required for consistency between the two observers' descriptions of, e.g., a rod passing through a ring). Note the resemblance to the Fitzgerald–Lorentz contraction of Chapter 1 — the same formula, but now derived as a necessary consequence of the postulates rather than invented to hide a null result.

Why transverse lengths cannot contract

The light-clock argument also rules out a transverse contraction. Let a light clock have mirror separation L_0 perpendicular to the relative motion. In the clock's rest frame its proper tick time is $\Delta t_0 = 2L_0/c$. Suppose, temporarily, that the separation perpendicular to the motion were kL_0 in a frame where the clock moves at speed v. In that frame,

$$\left(\frac{c\Delta t}{2}\right)^2 = \left(\frac{v\Delta t}{2}\right)^2 + (kL_0)^2. \tag{2.8}$$

Using the already established time-dilation result $\Delta t = \gamma\Delta t_0$ and $\Delta t_0 = 2L_0/c$ gives $k^2 = \gamma^2(1 - v^2/c^2) = 1$. Therefore $k = 1$: $y' = y$ and $z' = z$, as the Lorentz transformation states directly. Only the component parallel to the relative motion contracts.

> **Warning**
> Length contraction is a statement about a *measurement* — the positions of both ends of an object recorded simultaneously in the observer's frame — not a statement about what a fast-moving object would look like in an instantaneous photograph. A photograph also encodes differences in light-travel time from different parts of the object to the camera, and properly accounting for those shows that a rapidly moving object appears *rotated* rather than simply flattened (the Terrell–Penrose effect). "Measured length" and "visual appearance" are genuinely different questions, easy to conflate but with different answers.

Application: Bell's Spaceship Paradox

Length contraction raises a subtle question about rigid bodies that is worth working through carefully, since getting it wrong is a common source of confusion. Consider two identical spaceships, initially at rest a fixed distance L_0 apart in some frame S, connected by a taut string of exactly that length. At $t = 0$ (in S), both ships fire identical engines, executing identical acceleration profiles, so that at every later instant in S each ship has exactly the same velocity as the other. Does the string, stretched between them, break?

The naive answer is "no": since both ships always move identically, the distance between them, measured in S, never changes from L_0. But this overlooks what happens to the string. The string is a physical object with its own proper length, and *in the frame in which it is momentarily at rest*, the separation of its endpoints must grow, not stay fixed — because the string's own rest frame changes as it accelerates, and simultaneously requiring the ships to keep the *same* S-frame separation L_0 is a different condition from requiring their separation to stay constant *as measured by the string itself*. Since the ships' separation, measured in S,

stays exactly L_0 while the ships (and any observer moving with them) are increasingly length-contracted relative to S, the string's own rest-frame length would have to *increase* to keep up — equivalently, the distance between the ships, measured in the ships' own instantaneous rest frame, grows beyond L_0 as their common speed increases. A string of fixed material length cannot stretch to match this demand, and it snaps.

The paradox dissolves once it is clear that "the two ships have identical velocity at every instant of S-time" is a statement *specific to frame S*; it does not mean the ships are relatively at rest, nor does it mean the string, whose integrity depends on physics internal to its own rest frame, experiences no strain. This scenario (due to John Bell, 1976, who reported that it caused genuine disagreement even among professional physicists at CERN) is a useful check on intuition: length contraction is a real, physical effect with real, physical (and sometimes destructive) consequences for extended objects undergoing acceleration, not merely a bookkeeping artifact of coordinate choices.

2.2 The Lorentz Transformation and Causality

The Lorentz Transformation

Time dilation and length contraction are special cases of a general coordinate transformation between inertial frames, replacing the Galilean transformation of Chapter 1. For frame S' moving at velocity v along the common x-x' axis relative to frame S, with origins coinciding at $t = t' = 0$, the **Lorentz transformation** is

$$x' = \gamma(x - vt), \qquad y' = y, \qquad z' = z, \qquad t' = \gamma\left(t - \frac{vx}{c^2}\right). \tag{2.9}$$

The inverse transformation (from S' back to S) has the same form with $v \to -v$:

$$x = \gamma(x' + vt'), \qquad t = \gamma\left(t' + \frac{vx'}{c^2}\right). \tag{2.10}$$

Two features are worth noting. First, in the limit $v \ll c$, $\gamma \to 1$ and the vx/c^2 term becomes negligible, and the Lorentz transformation reduces to the Galilean transformation — special relativity does not discard Newtonian kinematics but contains it as the low-speed limit. Second, the time coordinate t' depends on *both* t and x: two events at different locations x that are simultaneous ($t_1 = t_2$) in frame S are, in general, *not* simultaneous in S' unless they also share the same x — a direct, quantitative statement of the relativity of simultaneity derived qualitatively above.

A quantity that has the same value in every inertial frame is called a **Lorentz invariant**. The most important one is the *spacetime interval* between two events,

$$(\Delta s)^2 = c^2(\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2, \tag{2.11}$$

which every inertial observer computes to be the same number, even though Δt and Δx individually differ between frames. This invariance can be verified directly by substituting the Lorentz transformation. The interval plays the same role for spacetime that ordinary Euclidean distance, $d^2 = \Delta x^2 + \Delta y^2$, plays for a plane: it is unchanged by a change of coordinates (there, a rotation of axes; here, a Lorentz "boost" to a new inertial frame), even though the individual coordinate differences are not.

Worked Example: Applying the Lorentz Transformation

> Multiplying a time by c converts it to a length ($c\,\Delta t$), which is why spacetime-interval and Lorentz-transformation calculations are often easiest with both space and time expressed in meters, as in the worked example below.

Two events occur in frame S: event 1 at $(x_1, t_1) = (0, 0)$, and event 2 at $(x_2, t_2) = (600\text{ m}, 1.00\ \mu\text{s})$. Frame S' moves at $v = 0.60c$ relative to S. Find the coordinates of both events in S', and verify that the interval is the same in both frames.

First, $\gamma = 1/\sqrt{1-0.36} = 1/\sqrt{0.64} = 1.25$. Event 1 is at the shared origin, so $(x_1', t_1') = (0, 0)$ trivially. For event 2,

$$x_2' = \gamma(x_2 - vt_2) = 1.25\left[600\text{ m} - (0.60)(3.00\times10^{8}\text{ m/s})(1.00\times10^{-6}\text{ s})\right] = 1.25(600\,\text{m} - 180\,\text{m}) = 525\,\text{m}, \tag{2.12}$$

$$t_2' = \gamma\left(t_2 - \frac{vx_2}{c^2}\right) = 1.25\left[1.00\times10^{-6}\text{ s} - \frac{(0.60)(600\text{ m})}{3.00\times10^{8}\text{ m/s}}\right] = 1.25(1.00\times10^{-6}\,\text{s} - 1.20\times10^{-6}\,\text{s}) = -0.25\ \mu\text{s}. \tag{2.13}$$

In S, the interval is $(\Delta s)^2 = c^2(\Delta t)^2 - (\Delta x)^2 = (300\text{ m})^2 - (600\text{ m})^2 = -270{,}000\text{ m}^2$ (using $c\,\Delta t = 300$ m). In S', $c\,\Delta t' = (3.00\times10^{8}\text{ m/s})(-0.25\times10^{-6}\text{ s}) = -75$ m, so $(\Delta s')^2 = (-75\text{ m})^2 - (525\text{ m})^2 = 5625\text{ m}^2 - 275{,}625\text{ m}^2 = -270{,}000\text{ m}^2$ — the same value, confirming invariance. Note also that $t_2' < 0$: event 2 occurs *before* event 1 in S', even though event 2 occurs after event 1 in S. This is only possible because, as the next section shows, the interval between these two events is spacelike.

Light Cones and Causality

Einstein's 1905 paper introduced the two postulates that resolved the experimental tension [Einstein, 1905b]; a later portrait of Einstein is included in Figure 2.4.

The Lorentz transformation permits the *order* of two events to reverse between frames, as the worked example above just demonstrated — but only for certain pairs of events, and understanding which pairs is essential to understanding why relativity does not undermine cause and effect.

> A handy number for quick estimates: light travels almost exactly 1 ft (about 0.30 m) in 1 ns, so a light-travel time in nanoseconds converts to a distance in feet at a glance.

Classify any two events by the sign of the interval between them, $(\Delta s)^2 = c^2(\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2$:

- **Timelike** separation, $(\Delta s)^2 > 0$: more time separates the events than would be needed for a signal traveling at or below c to get from one to the other. Equivalently, $|\Delta x|/|\Delta t| < c$. A material particle, which travels slower than c, can pass through both events, so one event can causally influence the other.
- **Spacelike** separation, $(\Delta s)^2 < 0$: the events are separated by more space than light could cross in the available time, $|\Delta x|/|\Delta t| > c$. No signal, of any kind consistent with the postulates, can connect them; neither event can be the cause of the other.



Figure 2.4: Albert Einstein in 1921. Photograph by Underwood & Underwood; public domain via Wikimedia Commons.

- **Lightlike** (or null) separation, $(\Delta s)^2 = 0$: the events lie exactly on each other's light cone, connectible only by a signal moving at precisely c.

Because $(\Delta s)^2$ is Lorentz-invariant, every inertial observer agrees on which of these three categories a given pair of events falls into, even though they may disagree on Δt and Δx individually. This is the resolution of the apparent puzzle in the worked example above: event 2 was found to occur *before* event 1 in S' despite occurring after it in S, but a direct calculation shows the interval between them, $-270{,}000\ \mathrm{m}^2$, is negative — the events are spacelike separated. Indeed, light would need $|\Delta x|/c = 2.0\ \mu$s to cross the 600 m separation, longer than the $\Delta t = 1.0\ \mu$s between the events. Because no signal could have traveled from event 1 to event 2 in the first place, no observer's disagreement about their time-ordering creates any physical contradiction: neither event could possibly have caused the other, in any frame.

The full set of events timelike-separated from a given event x and in its future is called x's **future light cone**; the analogous set in the past is its **past light cone**. Events in x's future light cone are exactly the events x could causally influence; events in its past light cone are exactly the events that could have influenced x. All observers agree on the light cone structure of any event, because it is built entirely from the invariant interval.

This structure is what protects causality in relativity. **If two events are timelike separated, every inertial observer agrees on their order** (the earlier one always is measured to occur first, in every frame) — only the *time interval* between them, not their order, is frame-dependent. Order can only reverse between frames, as in the worked example, for spacelike-separated events — and by construction, no causal signal can link spacelike-separated events anyway, since that would require traveling faster than c. If it were somehow possible to send a signal faster than c — a hypothetical "subspace radio" — that signal would connect two spacelike-separated events, and by the argument above there would exist a frame in which the signal is received *before* it is sent, an unambiguous violation of cause and effect. This is the deep reason the speed limit c is not merely an engineering inconvenience but is woven into the logical structure of the theory: **no particle, signal, or influence of any kind can propagate faster than c without permitting effects to precede their causes** in some valid inertial frame.

Velocity Addition

If a particle moves with velocity u'_x along the x'-axis of frame S', and S' moves at velocity v relative to S, what velocity u_x does the particle have in S? Differentiating the Lorentz transformation (or applying it to two nearby events along the particle's path) gives the **relativistic velocity-addition formula**:

$$u_x = \frac{u'_x + v}{1 + \dfrac{u'_x v}{c^2}}. \tag{2.14}$$

For $u'_x, v \ll c$, the denominator is nearly 1 and this reduces to the familiar Galilean rule $u_x \approx u'_x + v$. But for $u'_x = c$ (light propagating in the $+x'$ direction), the formula gives

$$u_x = \frac{c + v}{1 + v/c} = c, \tag{2.15}$$

for *any* $v < c$ — exactly consistent with the second postulate: adding any sub-light velocity to c still gives c. More generally, if $|u'_x| < c$ and $|v| < c$, then $|u_x| < c$ always: no combination of sub-light velocities, added relativistically, ever reaches or exceeds c.

Worked Example: Explaining the Fizeau Coefficient

Chapter 1 described Fizeau's 1851 measurement of light's speed in flowing water, which showed a puzzling *partial* drag coefficient $f = 1 - 1/n^2$ rather than full or zero entrainment. This is now straightforward to explain, with no separate assumption about the ether at all. Let S be the lab frame and S' the frame of the flowing water, moving at speed $v \ll c$ relative to the lab. In the water's own rest frame, light travels at the ordinary speed for that medium, $u'_x = c/n$. By the velocity-addition formula, the speed measured in the lab is

$$u_x = \frac{c/n + v}{1 + \dfrac{(c/n)v}{c^2}} = \frac{c/n + v}{1 + v/(nc)}. \tag{2.16}$$

Since $v \ll c$, expand the denominator using $(1 + x)^{-1} \approx 1 - x$ for small $x = v/(nc)$:

$$u_x \approx \left(\frac{c}{n} + v\right)\left(1 - \frac{v}{nc}\right) \approx \frac{c}{n} + v - \frac{v}{n^2} + O(v^2/c) = \frac{c}{n} + v\left(1 - \frac{1}{n^2}\right), \tag{2.17}$$

dropping the term of order v^2/c, which is negligible for the modest flow speeds (∼few m/s) used in the experiment. This is exactly Fizeau's measured result, with the drag coefficient $f = 1 - 1/n^2$ emerging directly from relativistic velocity addition applied to ordinary light propagation in a moving medium — no partial ether entrainment need be assumed at all.

> **See Also**
> The Need for Relativity works through Fizeau's original 1851 measurement and the ad hoc Fresnel drag coefficient it required under the ether hypothesis, along with the Michelson–Morley and Kennedy–Thorndike null results that motivated Einstein's postulates in the first place.

2.3 Doppler Effect, Spacetime Diagrams, and Magnetism

The Relativistic Doppler Effect

The classical Doppler effect — the pitch of an ambulance siren rising as it approaches and falling as it recedes — arises from the finite travel time of successive wave crests to a stationary observer. Light shows an analogous effect, but with a relativistic correction on top of it, because the *source's clock itself* runs slow according to the observer, an effect with no classical counterpart at all.

It is worth being precise about what the classical effect does and does not depend on, because the difference is the whole point. Figure 2.5 is the acoustic case, with the source and the observer independently movable. Move the source toward a stationary listener at speed u and the received frequency is $f_0/(1-u/v_s)$; leave the source alone and move the *listener* toward it at the same u and the answer is $f_0(1 + u/v_s)$ instead. The two disagree at second order in u/v_s, and they must: the air is a medium, the medium picks out a frame, and "which one is really moving" is a question sound can answer. Light has no such medium, so the formula derived below can depend on the relative velocity and on nothing else — and that constraint alone is nearly enough to fix it.

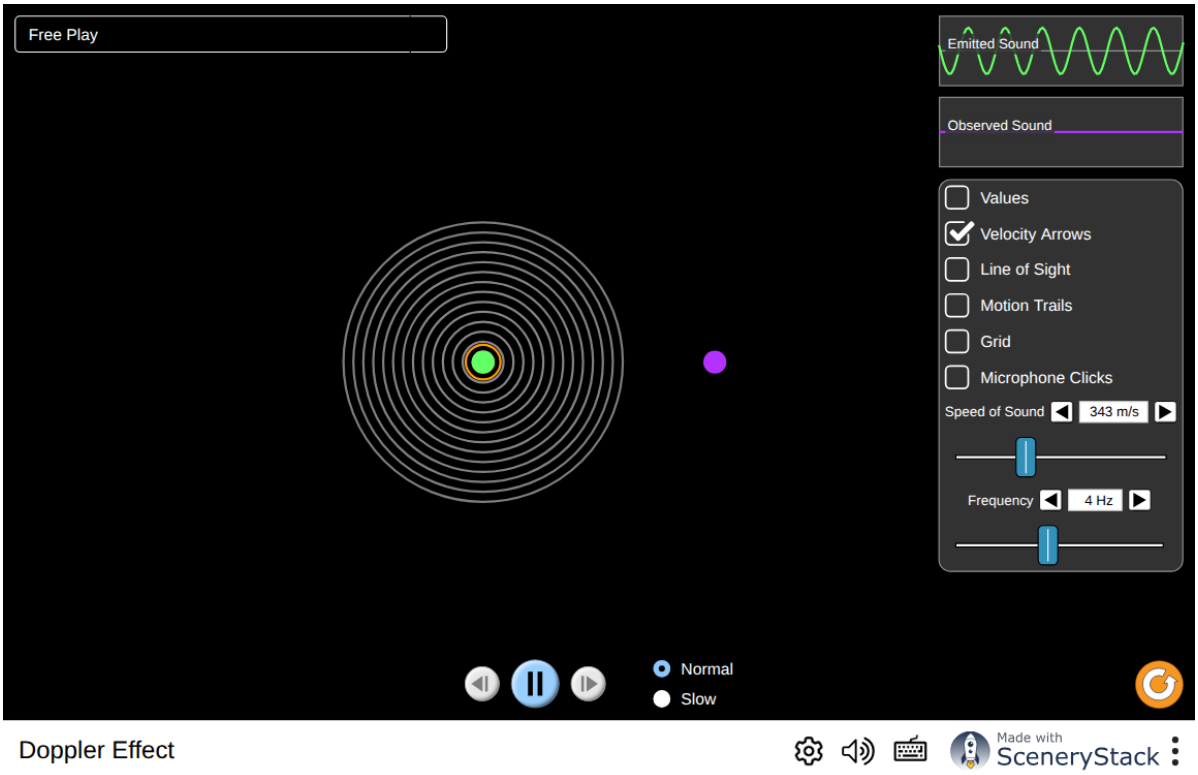

Figure 2.5: The *classical* Doppler effect, for sound in air. Drag the source and the observer independently and compare the shift produced by moving one against the shift produced by moving the other; the asymmetry between the two cases is the signature of a medium, and it is what disappears in the relativistic formula. Interactive simulation: Doppler Effect

Consider a source emitting light of proper frequency f_0 (the frequency measured by an observer at rest relative to the source), receding directly away from an observer at speed v. Two effects combine:

1. **Time dilation** of the source's own clock, as measured by the observer: successive wave crests are emitted at intervals of $\gamma\Delta t_0$ in the observer's frame, rather than $\Delta t_0 = 1/f_0$, purely because the moving source's clock runs slow.
2. **Light travel time**, which increases from one crest to the next because the source is receding, stretching the effective interval further by an additional factor.

Carrying out this calculation (Exercise 2.14 outlines the steps) gives the **relativistic Doppler formula** for a source receding directly away from the observer:

$$f_{\text{obs}} = f_0\sqrt{\frac{1 - v/c}{1 + v/c}}, \qquad \text{(receding)} \tag{2.18}$$

with the source's speed v measured relative to the observer; a source approaching directly gives the same formula with $v \to -v$, i.e.,

$$f_{\text{obs}} = f_0\sqrt{\frac{1 + v/c}{1 - v/c}}, \qquad \text{(approaching)}. \tag{2.19}$$

For $v \ll c$, expanding to first order in v/c recovers $f_{\text{obs}} \approx f_0(1 \mp v/c)$, the same leading-order shift familiar from the classical (sound) Doppler effect. The distinctively relativistic feature appears at *second* order and in a special geometric case: even a source moving **perpendicular** to the line of sight at the instant of emission — for which there is no classical Doppler shift at all, since the source is neither approaching nor receding — shows a purely relativistic **transverse Doppler shift**,

$$f_{\text{obs}} = \frac{f_0}{\gamma}, \tag{2.20}$$

a direct consequence of time dilation alone, with no light-travel-time contribution. This effect was confirmed with high precision by Ives and Stilwell (1938), who measured the frequency of light emitted by fast-moving hydrogen atoms viewed from the side, finding exactly the $1/\gamma$ shift predicted by time dilation and none of the shift a purely classical (source-frame) Doppler picture would have predicted.

The relativistic Doppler effect is not a laboratory curiosity. The cosmological redshift of light from distant galaxies is also a redshift, conventionally written $1 + z \equiv f_0/f_{\text{obs}}$, but it must be calculated using the expansion of spacetime rather than by applying this special-relativistic formula to arbitrarily distant galaxies. The same special-relativistic physics, applied to radio signals rather than light, is a correction that must be built into GPS satellite transmissions alongside the time-dilation correction already discussed.

Worked Example: Redshift from a Receding Galaxy

A distant galaxy's hydrogen emission line, with rest-frame (proper) wavelength $\lambda_0 = 656.3$ nm, is observed at $\lambda_{\text{obs}} = 682.0$ nm. Find the galaxy's recession speed.

Since $f = c/\lambda$, the frequency ratio is $f_{\text{obs}}/f_0 = \lambda_0/\lambda_{\text{obs}} = (656.3\text{ nm})/(682.0\text{ nm}) = 0.9623$. Setting this equal to $\sqrt{(1-v/c)/(1+v/c)}$ and squaring gives $0.9260 = (1-v/c)/(1+v/c)$. Solving for v/c:

$$0.9260(1+v/c) = 1 - v/c \implies v/c(1+0.9260) = 1 - 0.9260 \implies v/c = \frac{0.0740}{1.9260} = 0.0384, \tag{2.21}$$

so $v \approx 0.038c \approx 1.15 \times 10^7$ m/s, receding.

Spacetime Diagrams

A useful way to visualize these effects is a **spacetime diagram**: a plot with x on the horizontal axis and ct (rather than t, so both axes share units of length) on the vertical axis, drawn in a chosen frame S. A particle at rest at some fixed x traces a vertical line (its **worldline**); a light ray traces a line at 45°, since $x = ct$. An observer moving at speed v in frame S has a worldline tilted from vertical by an angle θ with $\tan\theta = v/c$.

> **Note**
> The spacetime diagram itself is due to Hermann Minkowski, one of Einstein's former mathematics professors, who recast special relativity in this geometric language in a famous 1908 address: "Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality." Minkowski's reformulation changed none of the 1905 theory's predictions, but it revealed the Lorentz transformation as nothing more exotic than a rotation of axes in a four-dimensional spacetime — a perspective that became indispensable for general relativity a decade later.

In frame S', that same moving observer's own x' and ct' axes are *not* perpendicular in the diagram as drawn in S: the ct' axis coincides with the observer's own worldline, while the x' axis — the locus of events simultaneous with the origin in S' — tilts up from the x-axis by the same angle θ that the ct' axis tilts from the ct-axis. This tilted-axis picture is a direct graphical statement of the relativity of simultaneity: the set of events an S'-observer calls "now" is not the same as the set an S-observer calls "now." Reading distances off a spacetime diagram requires care (the Lorentz-transformed axes are not orthogonal in the Euclidean sense), but the picture makes clear that simultaneity, not merely elapsed time, is the coordinate that differs between frames.

The light cone of the previous section appears naturally on such a diagram: the two 45° lines through any event x divide the diagram into the future (above both lines), the past (below both lines), and two "elsewhere" regions (spacelike-separated from x), to either side. As a frame's axes tilt with increasing relative speed v, they close in like a pair of scissors toward — but, for any $v < c$, never quite reaching — the 45° light-cone lines themselves. This is the geometrical statement that no continuous process of acceleration can bring a massive object's worldline from a slope less steep than 45° to one that equals or exceeds it: velocities do not simply add in relativity, and the speed c is a genuine asymptote, never attained.

The causal regions and their light-speed boundaries are shown in Figure 2.6.

The simulation in Figure 2.7 draws these diagrams live. Sliding the relative speed tilts the x' and ct' axes toward the light cone exactly as described above, and dragging an event shows

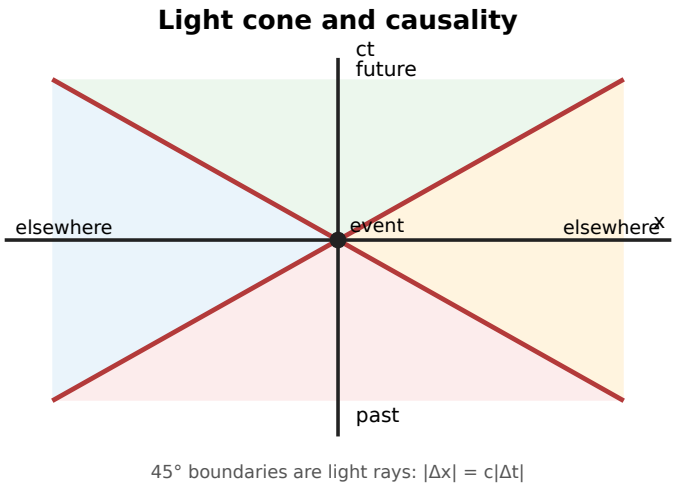

Figure 2.6: The light cone divides events into the causal future, causal past, and spacelike-separated "elsewhere." Its 45° boundaries are the worldlines of light in units where the horizontal coordinate is x and the vertical coordinate is ct. Original schematic by the author.

how its coordinates — and, on the other screens, the reading of a moving light clock and the aging of the traveling twin — change between frames.

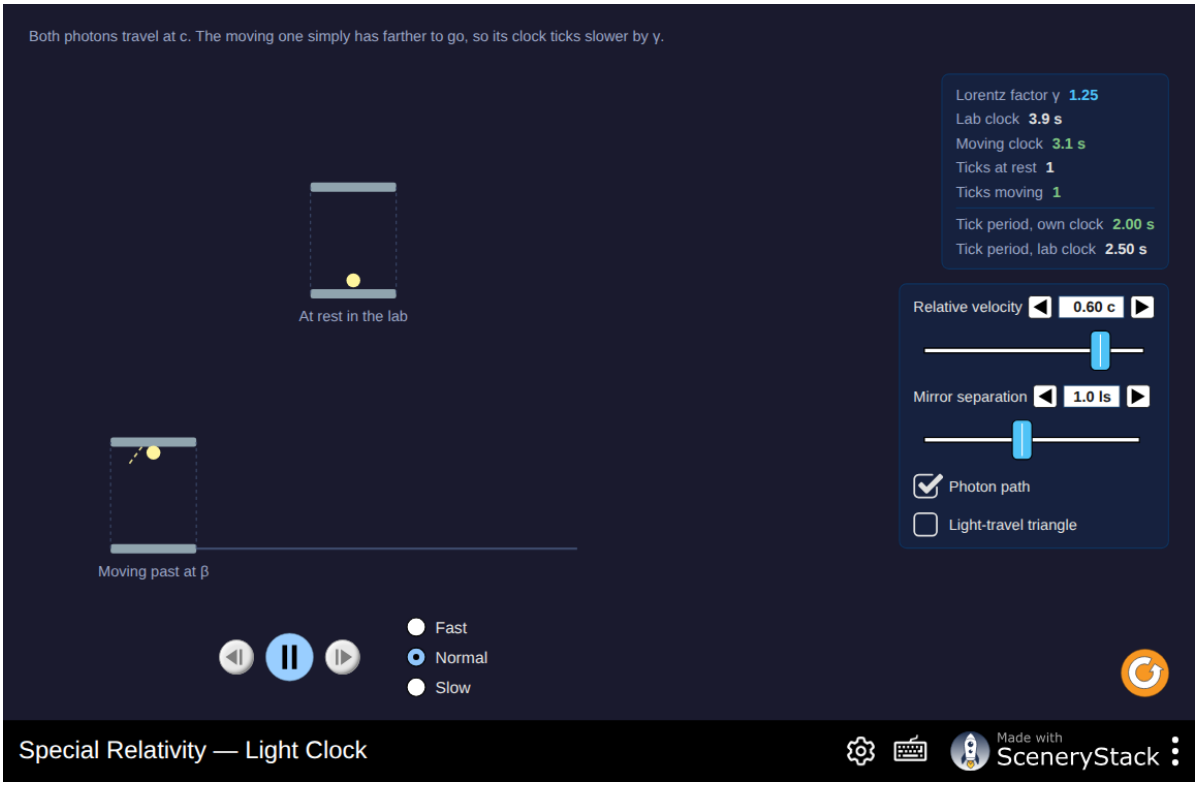

Figure 2.7: Special relativity in five screens: a moving light clock, an interactive Minkowski diagram, the ladder-and-barn paradox, the twin paradox, and the relativistic Doppler effect. Set the relative speed and watch the tilted axes close on the 45° light-cone lines without ever reaching them. Interactive simulation: Special Relativity

Application: Magnetism as a Relativistic Effect

It may seem that special relativity, having explained a handful of subtle high-precision optical experiments, is otherwise disconnected from everyday electromagnetism. In fact one of the most familiar phenomena in electromagnetism — magnetism itself — can be understood as a direct consequence of length contraction applied to electric charge.

Consider an idealized long wire with equal positive and negative linear charge densities moving in opposite directions along the wire, so that the wire's *net* charge density is zero in the lab frame. In that frame the wire carries a current. A test charge moving parallel to the wire can therefore experience a magnetic force, even though the wire produces no net electric field outside it. In the frame in which the two charge species move at equal and opposite speeds, their density changes by the same length-contraction factor, so their charge densities still cancel. If the test charge is moving in this frame, the magnetic field accounts for its force.

Now view the same situation from a frame in which the test charge is at rest (equivalently, an observer moving alongside it). By relativistic velocity addition, the two species of charge carriers in the wire no longer move at equal speeds relative to this observer—one is sped up and the other slowed down. Because length contraction depends on speed, their charge densities are no longer equal: the wire acquires a net linear charge density in this frame and therefore an electric field. That electric field produces the same physical deflection of the test charge that the magnetic field produced in the original frame; electric and magnetic fields have mixed under the change of frame.

Both descriptions—magnetic force in one frame and electric force from a net charge density in another—refer to the same underlying physical event, and both observers must agree on the test charge's worldline. Requiring this agreement forces the electric and magnetic fields to mix under a change of frame, consistently with the Lorentz transformation. In this setup, magnetism is the part of the electromagnetic interaction that appears when charges and the observer are in relative motion; electric and magnetic fields are not separate forces but frame-dependent components of one electromagnetic field. This unification is one of the clearest illustrations that relativity is not a remote, exotic correction confined to particle accelerators and GPS satellites, but a structural feature of the electromagnetic force that operates, imperceptibly, in every electric motor and every compass needle.

2.4 Summary

- Because both postulates must hold simultaneously, two events simultaneous in one inertial frame are, in general, not simultaneous in another: **simultaneity is relative**. Simultaneity within a single frame can be defined operationally by radar synchronization: bounce a light signal off a distant clock and assign it the round-trip midpoint time.
- A clock moving at speed v relative to an observer runs slow by the Lorentz factor $\gamma = 1/\sqrt{1 - v^2/c^2}$: $\Delta t = \gamma\,\Delta t_0$, where Δt_0 is the **proper time**, measured in the clock's own rest frame. This is **time dilation**, confirmed directly by observations such as cosmic-ray muon decay (Rossi–Hall, 1941).
- The **twin paradox** is resolved by recognizing that the traveling twin's worldline is not inertial throughout the trip (it includes acceleration at turnaround), breaking the naive symmetry between the twins; the twin who departs from and returns to a single inertial frame ages more.
- An object of proper length L_0 (measured in its own rest frame) has length $L = L_0/\gamma$



Figure 2.8: A frame-by-frame animated version of the length-contraction argument above, with the current-carrying wire drawn explicitly in both frames. Watch on YouTube: How Special Relativity Makes Magnets Work

as measured by an observer relative to whom it moves along its length. This is **length contraction**; only lengths along the direction of motion are affected.

- **Bell's spaceship paradox** shows that length contraction has real physical consequences: two spaceships accelerating identically, as measured in a common frame S, maintain constant separation in S, but a string connecting them experiences increasing strain and eventually breaks, because the string's own proper length would have to grow to keep pace with the ships' increasing length-contraction relative to S.
- The **Lorentz transformation**, $x' = \gamma(x - vt)$, $t' = \gamma(t - vx/c^2)$, replaces the Galilean transformation and reduces to it for $v \ll c$. The spacetime interval $(\Delta s)^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$ is the same in every inertial frame.
- Pairs of events are classified as **timelike** ($(\Delta s)^2 > 0$), **spacelike** ($(\Delta s)^2 < 0$), or **lightlike** ($(\Delta s)^2 = 0$) separated; all observers agree on this classification and on the time-order of timelike-separated events, which is what prevents faster-than-light signaling from creating causal paradoxes.
- Velocities combine via $u_x = (u'_x + v)/(1 + u'_x v/c^2)$, which reduces to Galilean addition at low speed, never produces a result at or above c when combining sub-light speeds, and reproduces the Fizeau drag coefficient $f = 1 - 1/n^2$ as its low-speed limit for light in a moving medium.
- The **relativistic Doppler effect**, $f_{\text{obs}} = f_0\sqrt{(1 \mp v/c)/(1 \pm v/c)}$ for direct recession/approach, combines the classical light-travel-time shift with time dilation of the source's clock; even purely transverse motion produces a shift, $f_{\text{obs}} = f_0/\gamma$, confirmed by the Ives–Stilwell experiment.
- Spacetime diagrams represent these effects graphically: a moving observer's axes of space and time are both tilted, by an equal angle, relative to a "stationary" observer's axes, and the light cone through any event marks the boundary of what that event can causally affect or be affected by.
- **Magnetism can be understood as a relativistic effect**: length contraction, applied unequally to the two oppositely-moving charge species in a current-carrying wire as seen from different frames, converts what looks like a purely magnetic force in one frame into a purely electric force (from a net charge density) in another.

2.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens by reading a spacetime diagram.

1. **Multiple choice, from a figure.** A second event lies in the region labeled "elsewhere," outside the origin's light cone. How is it separated from the origin? Figure description: Spacetime diagram showing the future and past light cones, timelike regions inside them, and spacelike regions outside them. Choices: (A) Spacelike: different inertial frames may reverse the event order, but no sub-light signal can connect the events. (B) Timelike: every inertial frame must agree that it occurs after the origin. (C) Lightlike: only a light signal can connect it to the origin. (D) Simultaneous in every inertial frame.

2. **True or false.** The proper time between two events is measured by one clock that is present at both events.

3. **Drag the words.** Complete the three central transformation rules. A moving clock accumulates _____ proper time, a moving object contracts _____, and all inertial observers calculate the same _____. The separate time and space differences are _____. (Terms: less; along the direction of motion; spacetime interval; frame dependent; more; in every direction.)

4. **Fill in the blanks.** Evaluate the Lorentz factor and length contraction at 0.8c. At v = 0.8c, γ = _____, so a moving object's length is _____ of its proper length.

5. **Mark the words.** Mark the two features that every inertial observer agrees on. "Observers agree on the spacetime interval and on an event pair's classification as timelike, spacelike, or lightlike; they need not agree on the separate distance, elapsed time, or simultaneity."

2.6 Conceptual Questions

1. Two clocks are synchronized in one frame. Why need observers moving relative to that frame not agree that they are synchronized?

2. Time dilation is reciprocal: each inertial observer says the other's clock runs slow. Explain why this is not a contradiction.

3. Why is length contraction stated only for the dimension parallel to relative motion?

4. A spaceship moves at $0.9c$. Why can neither ordinary addition nor any sequence of boosts make its measured speed exceed c?

5. What feature of a spacetime diagram distinguishes a possible causal influence from a spacelike separation?

6. Explain why the twin scenario is not symmetric once one twin turns around.

2.7 Problems

Exercise 2.1
(Straightforward)

A spaceship passes Earth at $v = 0.80c$. A clock on the ship ticks off exactly 1.00 s of proper time between two events at the same location on the ship. (a) What time interval between these two events is measured by an observer on Earth? (b) What is γ for this speed?

Exercise 2.2
(Straightforward)

A meter stick at rest in frame S' is oriented entirely along the direction of relative motion. If S' moves at $v = 0.60c$ relative to S, what length is the stick measured to have in S?

Exercise 2.3
(Straightforward)

Two events occur at the same time $t = 0$ in frame S, at positions $x_1 = 0$ and $x_2 = 300$ m. Frame S' moves at $v = 0.50c$ relative to S along the x-axis, with origins coinciding at $t = t' = 0$. Use the Lorentz transformation to find t'_1 and t'_2, and confirm that the two events are not simultaneous in S'.

Exercise 2.4
(Straightforward)

A cosmic-ray muon is created at an altitude of 15 km, moving straight down at $v = 0.998c$. Its proper mean lifetime is 2.2 μs. (a) Using time dilation, find the mean lifetime as measured in Earth's frame, and the mean distance the muon travels before decaying in that frame. (b) Working instead in the muon's rest frame, use length contraction to find the depth of the 15 km atmospheric layer as the muon measures it, and show your two calculations of whether the muon is likely to reach the ground agree.

Exercise 2.5
(Straightforward)

Two spaceships approach each other, each moving at speed $0.75c$ relative to Earth, in opposite directions. (a) What speed does an observer on one ship measure for the other ship, using relativistic velocity addition? (b) Explain why simply adding $0.75c + 0.75c$ would give an unphysical answer, and identify which postulate this would violate.

Exercise 2.6
(Moderate)

Sketch a spacetime diagram (axes x and ct) in the rest frame S of a laboratory. Draw the worldline of a particle at rest at $x = 2$ m, the worldline of a particle moving at $v = 0.5c$ starting from the origin, and the worldline of a light pulse emitted from the origin at $t = 0$. Identify the angle each worldline makes with the vertical axis.

Exercise 2.7
(Moderate)

Twin Bob leaves Earth at $v = 0.80c$, travels to a star 8.0 light-years away (in Earth's frame), immediately turns around, and returns at the same speed. (a) How much time elapses on Earth, according to Alice? (b) How much proper time elapses for Bob over the whole round trip? (c) Explain, referring specifically to Bob's turnaround, why it is Bob and not Alice who ages less, even though each twin sees the other's clock running slow during the constant-velocity legs of the trip.

Exercise 2.8
(Moderate)

Two events are separated by $\Delta x = 5.0 \times 10^8$ m and $\Delta t = 1.0$ s in frame S. (a) Determine whether the interval between them is timelike, spacelike, or lightlike. (b) If it is spacelike, explain why no observer's disagreement about which event happened first can create a causality paradox. (c) Find the speed v (as a fraction of c) of a frame S' in which the two events are simultaneous, if such a frame exists.

Exercise 2.9
(Moderate)

Light travels through a block of glass with refractive index $n = 1.50$ moving at $v = 20$ m/s relative to the lab, in the same direction as the light. Using the relativistic velocity-addition formula (not the approximation), find the light's speed in the lab frame, and compare the result to the Fizeau-formula approximation $c/n + v(1 - 1/n^2)$.

Exercise 2.10
(Moderate)

A wire carries equal and opposite densities of positive and negative charge moving at $\pm 0.90c$ relative to the lab frame (an exaggerated, but illustrative, speed), with the wire electrically neutral in the lab frame. (a) Qualitatively explain, using length contraction, why an observer moving alongside one species of charge carrier would measure a net charge density on the wire. (b) Explain why this observer must nonetheless agree with the lab-frame observer about whether a nearby test charge, initially at rest in the observer's frame, accelerates toward or away from the wire.

Exercise 2.11
(Moderate)

A spacecraft recedes directly from Earth at $v = 0.60c$, transmitting a radio signal at proper frequency $f_0 = 100.0$ MHz. (a) What frequency does a receiver on Earth measure? (b) If the same spacecraft instead approached Earth at the same speed, what frequency would be measured? (c) Explain why the approaching and receding frequencies are not simply related by $f_0(1 \pm v/c)$, i.e., why they are not symmetric about f_0 the way the low-speed (classical) approximation would suggest.

Exercise 2.12
(Challenging)

A star's light source moves in a circular orbit, at speed $v = 0.30c$, around a companion too dim to see, so that at one point in the orbit its velocity is purely transverse to the line of sight to Earth. (a) Find the fractional frequency shift, $(f_0 - f_{\text{obs}})/f_0$, expected at that instant from the transverse Doppler effect alone. (b) Explain why this shift is always a redshift (lower observed frequency), regardless of the direction of the transverse motion, unlike the ordinary (longitudinal) Doppler shift, which can be a redshift or blueshift depending on direction.

Exercise 2.13
(Challenging)

In Bell's spaceship paradox, both ships execute identical acceleration profiles as measured in frame S, so that their S-frame separation stays fixed at L_0. Explain why an observer riding along with either ship would nonetheless measure the *other* ship to be receding, at least for part of the trip, and connect this to the reason the connecting string comes under increasing tension. (You do not need to compute the tension quantitatively — a clear qualitative argument, referencing length contraction of the string versus the fixed S-frame separation of the ships, is sufficient.)

Exercise 2.14

(Challenging)

Outline the derivation of the relativistic Doppler formula for direct recession, by combining (a) the time-dilation factor γ for the interval between successive wave crests as emitted, and (b) the classical light-travel-time stretching factor $(1 + v/c)$ that arises because the source recedes an additional distance $v\Delta t$ between emitting successive crests. Show that multiplying these two factors and using $\Delta t_{\text{obs}} = \gamma(1 + v/c)\Delta t_0$ reproduces $f_{\text{obs}} = f_0\sqrt{(1 - v/c)/(1 + v/c)}$.

Chapter 3

Relativistic Dynamics

Learning Objectives

By the end of this chapter, you should be able to:

- Explain why the Newtonian definition of momentum, $p = mu$, is not conserved in all inertial frames once the Lorentz transformation replaces the Galilean transformation, and derive the corrected, relativistic definition from a symmetric elastic-collision thought experiment.
- Apply the relativistic expressions for momentum, kinetic energy, and total energy.
- Use the mass–energy relation $E = mc^2$ and the energy–momentum invariant $E^2 = (pc)^2 + (mc^2)^2$.
- Distinguish rest mass, relativistic momentum, kinetic energy, and total energy, and apply these to problems involving massive particles and photons.
- Represent energy and momentum as components of a single four-vector, and use conservation of this four-vector to analyze collisions, decays, and annihilation events.
- Transform to the center-of-momentum frame of a collision and compute threshold energies for particle-production reactions.
- Recognize the low-speed (Newtonian) and high-speed (ultrarelativistic) limits of the relativistic energy and momentum expressions.
- Explain why the combination of momentum blowing up as $u \to c$ and the causality argument of Chapter 2 together rule out faster-than-light material particles.

Introduction

Chapter 2 replaced the Galilean transformation with the Lorentz transformation, because only the latter is consistent with the invariance of the speed of light. But momentum and energy in Newtonian mechanics are defined and conserved using Galilean kinematics: $p = mu$, conserved because Newton's third law and Galilean-invariant forces guarantee it in every Galilean frame. Once the underlying kinematics changes, the old definitions of momentum and energy no longer transform consistently between inertial frames, and a collision that conserves Newtonian momentum in one frame will not, in general, conserve it in another frame related by a Lorentz transformation. This chapter derives the corrected definitions — relativistic momentum and relativistic energy — that *are* conserved in every inertial frame, assembles them into a single four-component object whose conservation law captures both at once, and works out their most important consequences: mass and energy are, up to a conversion factor c^2, the same quantity, and a whole new class of practical calculations — threshold energies for particle production —

becomes possible.

3.1 Relativistic Momentum

Deriving Relativistic Momentum

Rather than simply asserting the corrected formula, it is worth seeing exactly where it comes from, because the argument is a direct descendant of the light-clock argument used to derive time dilation in Chapter 2, and it makes clear that the correction is forced on us by the Lorentz transformation, not chosen for convenience.

Consider two identical particles, A and B, each of rest mass m, in an idealized elastic scattering event. Choose the incoming and outgoing velocities so that the x components are unchanged while each particle's y component reverses sign. In frame S, let B's transverse speed be w_0. Let S' move at speed v along the x-axis relative to S, and choose S' so that A has no x' motion. In S', let A's transverse speed be the same w_0. Because $\Delta y = \Delta y'$ while $\Delta t = \gamma(v)\left(\Delta t' + v\Delta x'/c^2\right) = \gamma(v)\Delta t'$ for A with $\Delta x' = 0$, the transverse speed measured in S is reduced by a factor of $\gamma(v)$:

$$u_y(A) = \frac{w_0}{\gamma(v)}, \qquad \text{as measured in } S. \tag{3.1}$$

In S, A therefore has an x component v and a y component $-w_0/\gamma(v)$ before the bounce, while B has a y component $+w_0$. After the bounce, both y components reverse sign. Writing the momentum of a particle moving at speed u as $p = m(u)\,u$, for some as-yet-undetermined function $m(u)$ that reduces to the particle's ordinary mass at $u = 0$, conservation of y momentum says that the incoming and outgoing transverse momenta are equal. Since the outgoing total is the negative of the incoming total, each must be zero; hence

$$m\big(u(A)\big)\,|u_y(A)| \;-\; m\big(u(B)\big)\,|u_y(B)| \;=\; 0, \tag{3.2}$$

where $u(A)$ and $u(B)$ denote the total (not just transverse) speeds of A and B in frame S. Substituting the magnitudes $|u_y(A)| = w_0/\gamma(v)$ and $|u_y(B)| = w_0$, and canceling the common factor of w_0,

$$\frac{m\big(u(A)\big)}{\gamma(v)} = m\big(u(B)\big). \tag{3.3}$$

Finally, take the limit $w_0 \to 0$: in this limit, B's transverse motion vanishes entirely, so $u(B) \to 0$ and $m(u(B)) \to m(0) \equiv m$, the particle's ordinary rest mass. Meanwhile A's transverse velocity component also vanishes, leaving A moving purely along x at the frame's relative velocity, so $u(A) \to v$. The boxed relation becomes

$$\frac{m(v)}{\gamma(v)} = m \quad\Longrightarrow\quad m(v) = \gamma(v)\,m, \tag{3.4}$$

for *arbitrary* v, since v was the (otherwise unconstrained) relative speed of the two frames. In other words, momentum conservation across inertial frames, together with the Lorentz transformation's effect on transverse velocities, forces the "effective mass" $m(u)$ appearing in $p = m(u)u$ to be exactly $\gamma(u)m$ — not simply m, as Newtonian mechanics assumed. This is the relativistic momentum derived below; unlike the Newtonian formula, it is exactly the definition needed to make momentum conservation frame-independent, consistent with the Lorentz transformation of Chapter 2.

Relativistic Momentum

The result of the argument above is

$$\vec{p} = \gamma m\vec{u} = \frac{m\vec{u}}{\sqrt{1 - u^2/c^2}}, \tag{3.5}$$

where m is the particle's **rest mass** — an intrinsic, frame-independent property of the particle, equal to the mass measured by an observer at rest relative to it — and u is the particle's speed in the frame in question. For $u \ll c$, $\gamma \to 1$ and this reduces to the Newtonian $\vec{p} = m\vec{u}$. As $u \to c$, however, $\gamma \to \infty$, so $p \to \infty$ as well: **a finite force acting for a finite time supplies only a finite impulse, whereas reaching the speed of light would require an unbounded impulse and work.** This is the precise dynamical reason no massive object can reach or exceed c, complementing the kinematic argument of Chapter 2 (relativistic velocity addition never produces $u \geq c$ from sub-light inputs).

> **Warning**
> Many older textbooks (and Einstein himself, in some early writing) call the combination γm the particle's "relativistic mass," treating $p = m_{\text{rel}}u$ as formally identical to the Newtonian formula with an ordinary, speed-dependent mass substituted in. Modern usage in physics avoids this language: m denotes only the single, frame-independent **rest mass**, and the factor of γ is understood to belong to the *kinematics* of momentum and energy, not to a literally growing mass. "Relativistic mass" also fails to generalize cleanly — it does not, for instance, correctly describe how a moving object's inertia to *transverse* versus *longitudinal* forces differs — so treat it as a historical notational convenience at best, not a physical mass increase.

Worked Example: Testing $p = \gamma mu$ Against Data

The formula $\vec{p} = \gamma m\vec{u}$ is not merely a theoretical nicety; it has been tested directly by measuring the momentum and speed of fast electrons and protons independently — momentum from the radius of curvature in a known magnetic field ($p = qBr$), speed from time-of-flight over a measured distance. Measurements of this kind confirm $p/(mu) = \gamma(u)$ across a wide range of speeds and rule out the Newtonian prediction $p/(mu) = 1$ at appreciable fractions of c. For a proton with $u = 0.60c$, for instance, $\gamma = 1/\sqrt{1 - 0.36} = 1.25$, so the relativistic momentum is 25% larger than the Newtonian formula predicts.

> **Tip**
> Before reaching for $\vec{p} = \gamma m\vec{u}$ or $K = (\gamma - 1)mc^2$ in a problem, check where the given speed sits relative to c. If $u/c \lesssim 0.1$ or so, $\gamma \approx 1 + \frac{1}{2}(u/c)^2$ is within about 0.5% of unity, and the Newtonian formulas $p \approx mu$ and $K \approx \frac{1}{2}mu^2$ are an excellent, much simpler approximation. If instead u is a substantial fraction of c — or if a kinetic energy K is given that turns out to be comparable to or larger than the rest energy mc^2 — the relativistic expressions are mandatory; the worked example at the end of this chapter shows just how badly the Newtonian formula fails once $K \gtrsim mc^2$.

3.2 Relativistic Energy

A parallel argument — demanding that the work-energy theorem, $dK = \vec{F} \cdot d\vec{x}$, hold with the relativistic force $\vec{F} = d\vec{p}/dt$ — leads to the relativistic kinetic energy

$$K = \gamma mc^2 - mc^2 = (\gamma - 1)mc^2. \tag{3.6}$$

Deriving $K = (\gamma - 1)mc^2$ from the work-energy theorem

Start from the definition of kinetic energy as the work done accelerating a particle from rest to speed u along a straight line:

$$K = \int F\,dx = \int \frac{dp}{dt}\,dx = \int \frac{dp}{dt}\,u\,dt = \int u\,dp, \tag{3.7}$$

using $dx = u\,dt$. With $p = \gamma mu = mu/\sqrt{1 - u^2/c^2}$, differentiating with respect to u gives

$$\frac{dp}{du} = m\left[\frac{1}{\sqrt{1 - u^2/c^2}} + \frac{u^2/c^2}{(1 - u^2/c^2)^{3/2}}\right] = \frac{m}{(1 - u^2/c^2)^{3/2}} = \gamma^3 m, \tag{3.8}$$

so $dp = \gamma^3 m\,du$, and

$$K = \int_0^u u'\,\gamma^3(u')m\,du' = m\int_0^u \frac{u'\,du'}{(1 - u'^2/c^2)^{3/2}}. \tag{3.9}$$

Substituting $s = 1 - u'^2/c^2$, $ds = -2u'\,du'/c^2$, the integral becomes elementary:

$$K = mc^2 \int_1^{1-u^2/c^2} \left(-\tfrac{1}{2}\right) s^{-3/2}\,ds = mc^2\Big[s^{-1/2}\Big]_1^{1-u^2/c^2} = mc^2\left(\frac{1}{\sqrt{1 - u^2/c^2}} - 1\right), \tag{3.10}$$

which is exactly $K = (\gamma - 1)mc^2$ — confirming, by direct integration of the relativistic work-energy theorem, the result quoted in the main text.

It is useful to expand this for $u \ll c$ using the binomial approximation $\gamma \approx 1 + \frac{1}{2}u^2/c^2 + \cdots$:

$$K \approx \left(1 + \frac{1}{2}\frac{u^2}{c^2}\right)mc^2 - mc^2 = \frac{1}{2}mu^2, \tag{3.11}$$

recovering the familiar Newtonian kinetic energy as the low-speed limit — a necessary consistency check, since Newtonian mechanics is extremely well tested at everyday speeds.

> The binomial approximation $(1 - x)^n \approx 1 + nx$ for small x, used here with $n = -1/2$ and $x = u^2/c^2$, recurs throughout relativity whenever a low-speed limit is needed; keeping the next term, $\gamma \approx 1 + \frac{1}{2}(u/c)^2 + \frac{3}{8}(u/c)^4$, is what quantifies exactly how good the Newtonian approximation is at a given speed.

The kinetic energy expression separates naturally into two terms: γmc^2, and a constant mc^2 subtracted off. Einstein's insight was to take both terms seriously as *energy*, not just their difference. Define the **total energy**

$$E = \gamma mc^2, \tag{3.12}$$

and the **rest energy**

$$E_0 = mc^2. \tag{3.13}$$

Then $K = E - E_0$: kinetic energy is the energy *above and beyond* the energy mc^2 a particle possesses simply by virtue of having rest mass m, even at rest ($u = 0$, $\gamma = 1$). This is the celebrated **mass–energy equivalence**: rest mass is a form of energy, convertible (in principle and, in nuclear and particle processes, routinely in practice) into other forms of energy, and vice versa. The conversion factor $c^2 \approx 9 \times 10^{16}$ m^2/s^2 is enormous, which is why converting even a small amount of rest mass releases a very large amount of energy — the physical basis of the energy released in nuclear fission and fusion, examined in Chapter 13.

Note

Einstein did not present $E = mc^2$ in his main June 1905 relativity paper; it appeared in a short follow-up note submitted that September, "Does the Inertia of a Body Depend Upon Its Energy Content?", which asked what happens to a body's mass when it emits energy as radiation. Einstein's argument there — analyzing a body emitting two equal light pulses in opposite directions, as measured in two different frames — concluded that a body losing energy L must lose mass L/c^2, and he suggested, remarkably prescient for 1905, that this might one day be tested with radioactive salts like those Marie and Pierre Curie were studying. The clean experimental confirmation, precisely measuring the mass deficit of a nucleus against the kinetic energy released in its reactions, would not come until nuclear physics matured decades later.

See Also

Nuclear Physics uses mass–energy equivalence to compute nuclear binding energies and the energy released in fission and fusion reactions — the direct, large-scale application of the $E = mc^2$ relation introduced here.

3.3 Four-Vectors and Threshold Energies

The Energy–Momentum Four-Vector

Momentum and energy are not independent; eliminating u and γ between $\vec{p} = \gamma m\vec{u}$ and $E = \gamma mc^2$ gives the **energy–momentum invariant**,

$$E^2 = (pc)^2 + (mc^2)^2, \tag{3.14}$$

a relation that holds for every particle, in every inertial frame.

Because pc has units of energy, particle physicists routinely quote momentum in units of MeV/c (or GeV/c) rather than kg · m/s — a momentum of 500 MeV/c means simply that multiplying by c gives an energy of 500 MeV. Watch for this unit throughout the worked examples and problems below.

This is directly analogous to the invariant spacetime interval $(\Delta s)^2 = c^2\Delta t^2 - \Delta x^2$ of Chapter 2, and the analogy is not superficial: just as (ct, x, y, z) can be assembled into a

single spacetime-displacement four-vector that transforms under a Lorentz boost according to the Lorentz transformation, the quadruple $(E/c, p_x, p_y, p_z)$ can be assembled into a single **energy–momentum four-vector**,

$$p^\mu = \left(\frac{E}{c},\, p_x,\, p_y,\, p_z\right), \tag{3.15}$$

which transforms from one inertial frame to another by exactly the same Lorentz transformation rule used for (ct, x, y, z) in Chapter 2 (with $ct \to E/c$ and $x \to p_x$). Its invariant magnitude is

$$\left(\frac{E}{c}\right)^2 - p_x^2 - p_y^2 - p_z^2 = (mc)^2, \tag{3.16}$$

which is precisely the energy–momentum relation in Equation (3.14), rearranged; the rest mass m plays the same role for the energy–momentum four-vector that the invariant interval plays for the spacetime-displacement four-vector — a quantity every observer computes to be the same, regardless of the frame in which E and $\vec{p}$ individually are measured. For a massive particle, the energy–momentum four-vector is proportional to its four-velocity. Its component ratio therefore gives the ordinary velocity, $v_x = p_x c^2/E$ (and similarly for the other components). A slower particle has a larger ratio of E/c to its momentum, as this expression requires.

The relation can be read geometrically as a right triangle, as shown in Figure 3.1: the total-energy term is the hypotenuse, while rest energy and momentum provide the two legs.

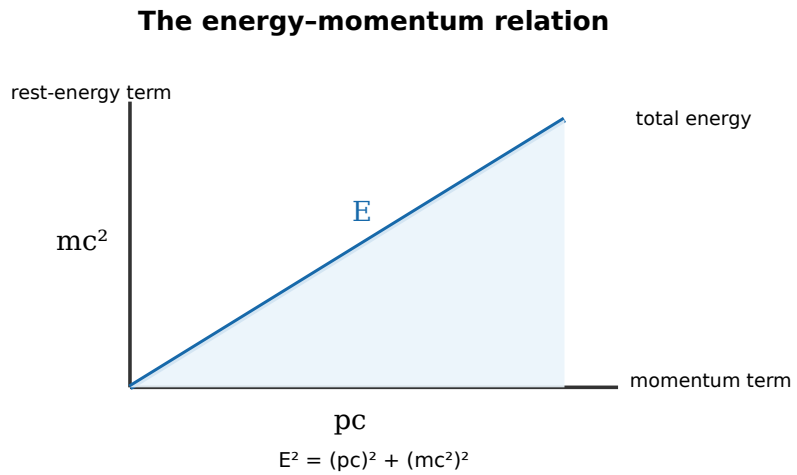

Figure 3.1: The energy–momentum relation as a right triangle: $E^2 = (pc)^2 + (mc^2)^2$. Original schematic by the author.

Three limits of the energy–momentum relation are worth committing to memory:

- **Massive particle at rest** ($p = 0$): $E = mc^2$, the rest energy alone.
- **Massless particle** ($m = 0$), such as a photon: $E = pc$. Massless particles carry momentum and energy but no rest energy, and — consistent with the momentum argument above, which forbids a *massive* particle from reaching $u = c$ — they travel at exactly c in every inertial frame.

- **Ultrarelativistic massive particle** ($pc \gg mc^2$, equivalently $K \gg mc^2$): $E \approx pc$ and $u \approx c$. The massless relation is therefore also an excellent approximation for a sufficiently energetic massive particle.

The energy–momentum relation is often more convenient than working with u and γ directly, particularly for high-energy particles and for photons, where speed is fixed at c and carries no information about energy.

Why the Four-Vector Formalism Earns Its Keep

The payoff of packaging $(E/c, \vec{p})$ as a single object is that **the energy–momentum four-vector is exactly conserved in every collision or decay, in every inertial frame, component by component** — precisely because the underlying, separately-conserved quantities (in every frame) are energy and the three components of momentum. This means a conservation calculation can be carried out entirely by four-vector addition: add up the four-vectors of everything going into a collision, add up the four-vectors of everything coming out, and set the two sums equal, four components at a time. Because the four-vector's magnitude, $(E/c)^2 - p^2 = (mc)^2$, is invariant, this magnitude can be computed in *whichever* frame is most convenient — often a frame in which one particle is initially at rest, or the frame in which the total three-momentum is zero (the subject of the next section) — and the resulting relation between energies and masses will hold in every other frame as well.

Electron–positron annihilation, revisited: consider an electron and a positron (each of mass m), both essentially at rest, annihilating. Charge conservation alone would permit $e^- + e^+ \to \gamma$ (a single photon), but the four-vector of the initial state is $(2mc, 0, 0, 0)$ (both particles at rest, energies mc^2 each, zero total momentum), while any single photon's four-vector must satisfy $E = pc$, i.e., have equal, nonzero energy and momentum magnitude — it cannot have zero momentum unless its energy is also zero. A single outgoing photon is therefore impossible; conservation of the energy–momentum four-vector, not merely of energy or momentum separately, forces at least two photons, emitted back-to-back so that their momenta cancel, each carrying energy mc^2. This is the same conclusion reached by separate energy and momentum arguments in Exercise 3.3 below, but the four-vector language makes clear that both conservation laws are really a single, unified statement, and this is exactly the physical process (positron annihilation, producing two back-to-back 511 keV gamma rays) exploited in medical positron-emission tomography (PET) scans to locate a radioactive tracer inside the body.

Center-of-Momentum Frame and Threshold Energies

> The CM frame is sometimes called the "center-of-mass frame," a holdover from Newtonian mechanics; in relativity, "center-of-momentum" is the more accurate name, since it is defined by zero total momentum, not by any geometric mass-weighted position.

Many practical problems in nuclear and particle physics — will one particle collision produce a new particle, or not? — are most easily solved by transforming to the **center-of-momentum (CM) frame**: the unique inertial frame in which the total three-momentum of a system is zero. Because $(E/c)^2 - p^2$ is Lorentz-invariant, the *total* invariant mass of a system of particles,

$$(Mc^2)^2 \equiv E_{\text{total}}^2 - (p_{\text{total}}c)^2, \tag{3.17}$$

is the same number whether computed in the lab frame or in the CM frame — but in the CM frame, where $p_{\text{total}} = 0$ by definition, it simplifies to $Mc^2 = E_{\text{total,CM}}$: the total CM-frame energy alone. This invariant M determines whether a reaction is kinematically possible: by conservation of the energy-momentum four-vector, a reaction that produces a set of final-state particles with total rest mass $\sum m_f$ is only possible if the *available* energy in the CM frame is at least $\left(\sum m_f\right) c^2$ — i.e., if $M \geq \sum m_f$. Other conservation laws and interaction probabilities can still forbid or suppress a reaction. The **threshold** condition is $M = \sum m_f$ exactly, corresponding, for massive final-state particles, to all of them being at rest relative to the CM frame (and hence relative to each other), with no leftover kinetic energy to spare.

The CM frame is not a relativistic invention, and it is worth recovering the non-relativistic intuition before leaning on it. Figure 3.2 runs elastic and inelastic collisions in one and two dimensions, with the center of mass drawn on the screen and a momentum diagram beside it. Whatever the pucks do, that marker glides on at constant velocity — the collision cannot touch it, because the internal forces cancel in pairs — which is what makes its rest frame a natural place to do the bookkeeping. Choose the masses and velocities so that the marker stands still, and the momentum diagram shows what has been bought: two arrows equal and opposite before the collision, two arrows equal and opposite after it, however much kinetic energy was lost in between. The relativistic version below keeps that structure exactly, replacing $m\vec{u}$ by $\gamma m\vec{u}$ and the total mass by the invariant M; what changes is that M is no longer the sum of the parts.

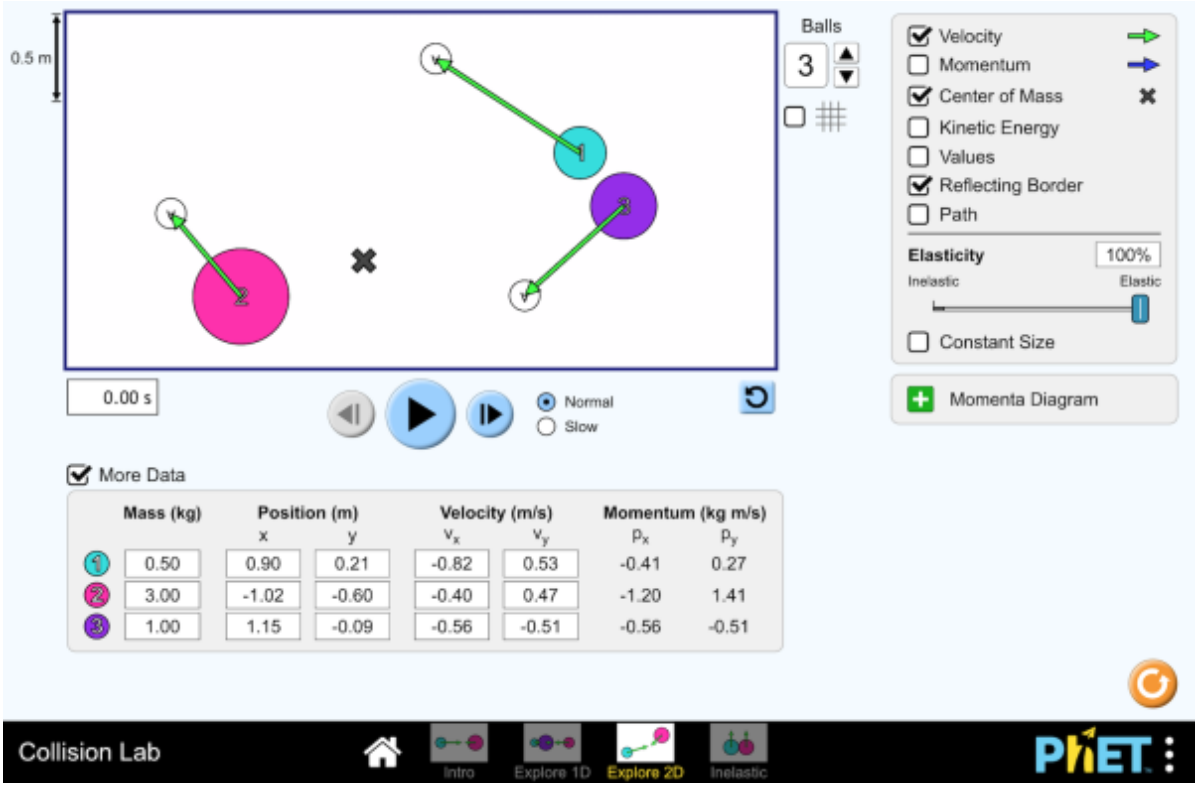

Figure 3.2: Classical collisions with the center of mass and the momentum vectors displayed. The center of mass moves at a velocity no collision can change, elastic or not, which is what makes its rest frame a natural place to do the bookkeeping — and what carries over, with $\gamma m\vec{u}$ in place of $m\vec{u}$, to the relativistic threshold calculations of this section. Interactive simulation: Collision Lab

Worked Example: The Threshold for Antiproton Production

In 1955, the Bevatron at Berkeley was built specifically to search for the antiproton, via the reaction

$$p + p \to p + p + p + \bar{p}, \tag{3.18}$$

a proton beam striking a stationary proton (in a hydrogen target), producing an additional proton–antiproton pair. (Baryon number and charge are each automatically conserved by this reaction; nothing prevents it kinematically once enough energy is available.) What is the minimum, or *threshold*, kinetic energy the beam proton must have, in the lab frame, where the target proton is at rest?

Let m be the proton (and antiproton) rest mass, $m_1 = m_2 = m$ for beam and target, and let $M = 4m$ be the total rest mass of the four final-state particles, all momentarily at rest in the CM frame at threshold. The invariant $M^2c^4 = E_{\text{total}}^2 - (p_{\text{total}}c)^2$ can be computed in the lab frame, where the target is at rest ($E_2 = mc^2$, $p_2 = 0$) and the beam proton has energy E_1 and momentum p_1:

$$(4mc^2)^2 = (E_1 + mc^2)^2 - (p_1c)^2 = E_1^2 + 2E_1mc^2 + m^2c^4 - (p_1c)^2. \tag{3.19}$$

Using $E_1^2 - (p_1c)^2 = (mc^2)^2$ (the beam proton's own invariant), this simplifies to

$$16m^2c^4 = m^2c^4 + 2E_1mc^2 + m^2c^4 \quad \Longrightarrow \quad E_1 = 7mc^2. \tag{3.20}$$

The threshold *kinetic* energy is $K_1 = E_1 - mc^2 = 6mc^2$. With $mc^2 = 938$ MeV for the proton, this is $K_1 = 6(938\text{ MeV}) = 5.6$ GeV — dramatically larger than the naive $2mc^2 = 2(938\text{ MeV}) = 1.9$ GeV one might have guessed from simply counting the rest-mass energy of the new particle pair. The extra factor of three arises because, in the lab frame, the newly created particles must all share the *same* velocity as the CM frame itself (since at threshold they are at rest *in* the CM frame, which is itself moving relative to the lab), so a substantial fraction of the beam's kinetic energy is unavoidably "wasted" maintaining the overall forward motion of the collision products rather than being converted into new rest mass. The Bevatron was deliberately designed to reach a beam energy of 6.2 GeV, comfortably above this threshold, and the antiproton was discovered there later that same year by Owen Chamberlain, Emilio Segrè, and collaborators.

This "wasted energy" problem is exactly why modern particle physics favors **colliders**, in which two beams travel toward each other and collide head-on. If the beam and target in the reaction above were replaced by two protons of equal and opposite momentum (so that the lab frame *is* the CM frame), the threshold condition becomes simply $2E_1 = 4mc^2$, i.e., $K_1 = mc^2$ per beam — six times less kinetic energy required per proton than the fixed-target case, precisely because no energy needs to be spent maintaining a net forward CM velocity.

The contrast between the two arrangements is summarized in Figure 3.3.

The historical setting of this threshold calculation is shown in Figure 3.4: the Bevatron was built in the 1950s to reach the energies needed to discover the antiproton.

Aside: Why There Are No Faster-Than-Light Massive Particles

Chapter 2 argued, from causality alone, that no signal or influence can travel faster than c without permitting effects to precede their causes in some valid inertial frame. This chapter's momentum formula, $p = \gamma mu$, gives an independent, purely dynamical reason a *massive*

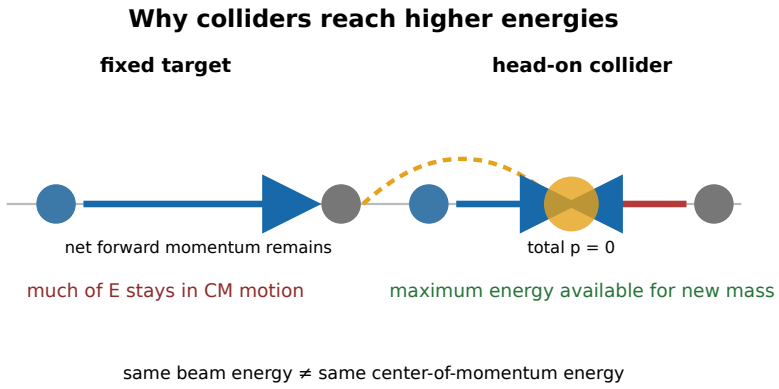

Figure 3.3: Fixed-target versus collider kinematics. In a head-on collider the center-of-momentum frame can coincide with the laboratory, so more of the beam energy is available to create rest mass. Original schematic by the author.



Figure 3.4: Interior of the former Bevatron building at Lawrence Berkeley National Laboratory. Photograph by Daniel Parks, 2010; CC BY 2.0 via Wikimedia Commons. The photograph shows the surviving facility structure, not the operating 1955 machine.

particle in particular can never reach or exceed c: as $u \to c^-$, $\gamma \to \infty$, so accelerating a massive particle arbitrarily close to c requires arbitrarily large — and, at $u = c$ itself, literally infinite — momentum and energy. No finite amount of work can supply this, so c is a strict, unreachable asymptote for any object with $m > 0$, approached but never attained no matter how long or how powerfully it is accelerated.

It is sometimes asked whether a hypothetical particle might simply be *born* moving faster than c, without ever having to accelerate through c — such a hypothetical particle is called a **tachyon**. Formally, applying the energy–momentum relation at $u > c$ gives a spacelike four-momentum, so the invariant mass-squared would be negative (often described informally as an "imaginary rest mass"). This is not the rest mass of an ordinary particle, and no such particle has been observed. More importantly, a tachyon would connect spacelike-separated events (Chapter 2) — and the causality argument there shows that *some* inertial observer would measure any spacelike-connecting signal to travel backward in time, arriving before it was sent. Thus the mass–energy relation excludes faster-than-light motion for ordinary massive particles, while the light-cone argument gives a separate causal reason that controllable faster-than-light signals cannot exist.

Worked Example: An Electron Accelerated Through a Potential Difference

An electron (rest energy $mc^2 = 0.511$ MeV) is accelerated from rest through a potential difference of 2.00 MV, gaining kinetic energy $K = qV = 2.00$ MeV.

Total energy: $E = K + mc^2 = 2.00 \text{ MeV} + 0.511 \text{ MeV} = 2.511 \text{ MeV}$.

Momentum: from $E^2 = (pc)^2 + (mc^2)^2$,

$$pc = \sqrt{E^2 - (mc^2)^2} = \sqrt{(2.511 \text{ MeV})^2 - (0.511 \text{ MeV})^2} = 2.458 \text{ MeV}, \tag{3.21}$$

so $p = 2.458 \text{ MeV}/c$.

Speed: from $E = \gamma mc^2$, $\gamma = E/mc^2 = (2.511 \text{ MeV})/(0.511 \text{ MeV}) = 4.914$, and $u = c\sqrt{1 - 1/\gamma^2} = 0.979c$.

Note that a Newtonian calculation of the speed from $K = \frac{1}{2}mu^2$ would give $u = c\sqrt{2K/mc^2} = c\sqrt{2(2.00 \text{ MeV})/(0.511 \text{ MeV})} \approx 2.8c$ — an unphysical result exceeding c, and a sharp reminder that the Newtonian kinetic-energy formula must not be used once K is comparable to or larger than mc^2.

3.4 Summary

- Because momentum must be conserved in every inertial frame under the Lorentz transformation, the correct definition is $\vec{p} = \gamma m\vec{u}$, not the Newtonian $m\vec{u}$. This can be derived by analyzing a symmetric elastic collision in two frames related by a boost: requiring conservation of transverse momentum, together with the Lorentz-transformation rule for transverse velocities, forces a moving particle's effective inertia to scale as $\gamma(u)m$.
- As $u \to c$, $p \to \infty$, which is why no massive particle can reach the speed of light; this is confirmed directly by measurements of $p/(mu)$ for fast electrons and protons, which match $\gamma(u)$ and rule out the Newtonian prediction of 1.
- Relativistic kinetic energy is $K = (\gamma - 1)mc^2$, which reduces to $\frac{1}{2}mu^2$ for $u \ll c$.
- **Total energy** $E = \gamma mc^2$ and **rest energy** $E_0 = mc^2$ satisfy $K = E - E_0$: mass is a form of energy (mass–energy equivalence), convertible to and from other forms of energy.

- The **energy–momentum invariant** $E^2 = (pc)^2 + (mc^2)^2$ holds for all particles in all frames; it reduces to $E = mc^2$ for a particle at rest and to $E = pc$ for a massless particle such as a photon, which necessarily travels at c.
- For an ultrarelativistic massive particle with $pc \gg mc^2$, the same invariant gives $E \approx pc$ and $u \approx c$.
- Energy and momentum combine into a single **energy–momentum four-vector**, $p^\mu = (E/c, p_x, p_y, p_z)$, which transforms like (ct, x, y, z) under a Lorentz boost and whose invariant magnitude is the rest mass, $(E/c)^2 - p^2 = (mc)^2$. Conservation of energy and momentum together is equivalent to conservation of this single four-vector.
- The **center-of-momentum (CM) frame** is the frame in which total momentum is zero; the invariant total mass M of a system, computed from $M^2c^4 = E_{\text{total}}^2 - (p_{\text{total}}c)^2$, sets the **threshold** condition, $M = \sum m_f$, for whether a reaction producing a given set of final-state particles is energetically possible, and fixed-target reactions require substantially more beam kinetic energy than head-on collider reactions to reach the same threshold.
- Combining the divergence of $p = \gamma mu$ as $u \to c$ with the causality argument of Chapter 2 rules out faster-than-light motion for ordinary massive particles and shows why controllable faster-than-light signals would create causal paradoxes.
- Newtonian expressions for momentum and kinetic energy are the $u \ll c$ (equivalently $K \ll mc^2$) limit of the relativistic expressions and must not be used when this condition fails.

3.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens on the energy–momentum diagram.

1. **Multiple choice, from a figure.** For a massless particle, what happens to the energy–momentum triangle shown above? Figure description: Right triangle with legs pc and mc squared and hypotenuse E, illustrating the invariant energy–momentum relation. Choices: (A) The rest-energy leg becomes zero, so the relation reduces to E = pc. (B) The momentum leg becomes zero, so the relation reduces to E = mc^2. (C) All three sides remain nonzero and equal. (D) The total-energy side becomes zero while momentum remains nonzero.

2. **True or false.** For the same total beam energy, a fixed-target experiment makes as much center-of-momentum energy available for new particles as a head-on collider.

3. **Drag the words.** Complete the relativistic momentum and energy relations. A massive particle has momentum p = _____, kinetic energy K = _____, and total energy E = _____. A massless photon obeys E = _____. (Terms: γmu; $(\gamma - 1)mc^2$; γmc^2; pc; mu; mc^2.)

4. **Fill in the blanks.** Evaluate the energy of a particle moving at 0.6c. At u = 0.6c, γ = _____, so the kinetic energy is _____ mc^2.

5. **Mark the words.** Mark the frame and system property used most directly in a threshold calculation. “In the center-of-momentum frame the total momentum is zero, and the system’s invariant mass determines which final-state rest masses can be created.”

3.6 Conceptual Questions

1. Why does applying a constant force fail to accelerate a massive particle through the speed of light?

2. Distinguish total energy, rest energy, and kinetic energy for a relativistic particle.

3. Why is rest mass invariant even though energy and momentum depend on the observer?

4. In a two-particle collision, why is the center-of-momentum frame especially useful for threshold calculations?

5. A photon has zero rest mass but nonzero momentum. Explain why this is consistent with the energy–momentum relation.

6. Why must momentum as well as energy be conserved when predicting whether a reaction is possible?

3.7 Problems

Exercise 3.1
(Straightforward)

A proton (rest energy 938 MeV) moves at $u = 0.900c$. Find (a) γ, (b) its total energy E, (c) its kinetic energy K, and (d) its momentum p (in MeV/c).

Exercise 3.2
(Straightforward)

Find the speed at which a particle's relativistic kinetic energy is 10% larger than the Newtonian prediction $\frac{1}{2}mu^2$. (Hint: compute $K_{\text{rel}}/K_{\text{Newt}}$ as a function of u/c and solve numerically or by successive approximation.)

Exercise 3.3
(Straightforward)

In electron–positron annihilation, an electron and a positron (each of rest energy 0.511 MeV), both essentially at rest, annihilate into two photons. (a) Use conservation of energy to find the energy of each photon, assuming they emerge with equal energies. (b) Use conservation of momentum to explain why the two photons must travel in exactly opposite directions.

Exercise 3.4
(Straightforward)

A pion at rest (rest energy 139.6 MeV) decays into a muon (rest energy 105.7 MeV) and a muon neutrino (treat as massless). Using conservation of energy and momentum together with $E^2 = (pc)^2 + (mc^2)^2$, find the kinetic energy of the muon produced.

Exercise 3.5
(Moderate)

Show algebraically that $E^2 = (pc)^2 + (mc^2)^2$ follows from $\vec{p} = \gamma m\vec{u}$ and $E = \gamma mc^2$ by eliminating u (use $\gamma^2(1 - u^2/c^2) = 1$).

Exercise 3.6
(Moderate)

The Sun radiates energy at a rate of about 3.8×10^{26} W. Use mass–energy equivalence to estimate the rate, in kg/s, at which the Sun is losing mass. Compare this rate, sustained for 4.6 billion years, to the Sun's mass of about 2.0×10^{30} kg, and comment on whether this loss is significant over the Sun's lifetime so far.

Exercise 3.7
(Moderate)

Fill in the missing step in the derivation of relativistic momentum: starting from the requirement $m(u(A))\,u_y(A) = m(u(B))\,w_0$ and the transverse-velocity relation $u_y(A) = w_0/\gamma(v)$, verify that taking $w_0 \to 0$ gives $u(A) \to v$ and $u(B) \to 0$, and hence that $m(v) = \gamma(v)\,m$.

Exercise 3.8
(Moderate)

Two identical lumps of putty, each of rest mass m and speed $u = 0.60c$ (in the lab frame), collide head-on and stick together, forming a single composite lump at rest in the lab frame. (a) Using conservation of the energy–momentum four-vector, find the rest mass M of the resulting composite lump, in terms of m. (b) Explain why $M \neq 2m$, and identify what has happened to the "missing" (or "extra") mass–energy.

Exercise 3.9
(Moderate)

A physicist wants to create a hypothetical new particle X, of rest mass $m_X = 10m_p$ (ten proton masses), via the fixed-target reaction $p + p \to p + p + X$, firing a proton beam at a stationary proton target. (a) Find the threshold beam kinetic energy, in units of $m_p c^2$, using the method of the worked example. (b) Find the threshold beam kinetic energy per proton if, instead, two beams of equal and opposite momentum collide head-on. (c) Compute the ratio of your two answers, and comment on why collider experiments are favored for producing very massive particles.

Exercise 3.10
(Challenging)

In an idealized absorption reaction, a photon of energy E_γ collides head-on with a stationary electron of rest mass m and is absorbed, forming a single recoiling particle of rest mass M. Use four-vector conservation to express M in terms of m, E_γ, and c. What is the particle's momentum in the lab frame?

Exercise 3.11
(Challenging)

A tachyon is hypothesized to have real, finite energy and momentum despite traveling at $u = 1.5c$. (a) Show that the relation $E^2 = (pc)^2 + (mc^2)^2$ can only be satisfied for $u > c$ if $m^2 < 0$. (b) Explain, using the light-cone argument of Chapter 2, why even setting aside the issue in (a), a signal carried by such a particle would create a causality paradox for some inertial observer.

Exercise 3.12

(Challenging)

Verify the claim in the worked example on antiproton production that a naive estimate of $2mc^2$ (simply the rest-energy cost of the new proton–antiproton pair) undercounts the true threshold kinetic energy of $6mc^2$. Do this by computing the speed of the center-of-momentum frame at threshold (i.e., the velocity of the frame in which the total momentum is zero) as a fraction of c, and explain qualitatively, in terms of this CM-frame velocity, where the "extra" energy beyond $2mc^2$ goes.

Part II

Wave Optics

Chapter 4

Interference of Light

Learning Objectives

By the end of this chapter, you should be able to:

- Describe light as an electromagnetic wave, relate its wavelength, frequency, and speed, and explain why detectors measure time-averaged intensity rather than the instantaneous field.
- Explain why interference and diffraction are invisible in everyday life, using the dimensionless ratio λ/a and the correspondence between wave optics and ray optics.
- State Huygens's principle and use it to explain how a wavefront propagates and spreads after an aperture.
- Convert a path length difference into a phase difference, and use the result to decide whether two overlapping waves interfere constructively or destructively.
- Derive the path-difference conditions for Young's double-slit experiment, and calculate fringe angles, fringe positions, fringe spacing, and the highest observable order.
- Explain why two independent light sources do not produce a stable interference pattern; distinguish temporal from spatial coherence, and estimate coherence length from a source's spectral width.
- Derive the two-slit intensity pattern by phasor addition, show that the central maximum is four times a single slit's intensity, and explain why this does not violate energy conservation.
- Generalize the phasor method to N equally spaced, idealized slits, and explain why the principal maxima sharpen as $1/N$ and brighten as N^2.
- Apply the thin-film interference conditions, correctly counting the π phase shift on reflection from a higher-index medium, to soap films, antireflection coatings, air wedges, and Newton's rings.
- Describe the operation of the Michelson interferometer, convert a mirror displacement into a fringe count, and connect the instrument to the Michelson–Morley experiment of Chapter 1.

Introduction

Chapters 1–3 built special relativity from a single empirical fact: every inertial observer measures the same speed c for light. That argument treated light purely as a *signal*, without asking what, physically, is doing the propagating. Maxwell's equations answer that question: light is an electromagnetic wave — a self-sustaining, traveling disturbance of the electric and magnetic

fields. If that is right, then light must do what every other wave does. It must **interfere**: two overlapping light waves must be able to reinforce each other in some places and cancel each other in others. And it must **diffract**: it must bend around obstacles and spread out after passing through an aperture, instead of casting perfectly sharp geometric shadows.

Neither effect is part of ordinary experience. Shadows do look sharp. Light does seem to travel in straight lines. Two flashlights shone on the same wall produce a brighter patch, not a striped one. For most of the seventeenth and eighteenth centuries this was taken as decisive: Newton championed a *corpuscular* theory, in which light is a stream of tiny particles, and his enormous authority kept the wave theory of his contemporary Christiaan Huygens in the minority for over a century. A large part of this chapter's job is to explain why the wave effects hide so effectively — the answer, worked out below, is that the wavelength of visible light is a few hundred nanometers, some three orders of magnitude below the smallest object you handle in a day and six or more below a doorway — and then to show what happens when you shrink the apparatus until they can no longer hide.

Thomas Young did exactly that in 1801. He passed light from a single source through two closely spaced slits and found, on a distant screen, not two bright bands but a whole series of alternating bright and dark fringes. There is no way to get dark bands by adding light to light in a particle picture: two streams of corpuscles cannot cancel. In a wave picture the explanation is immediate, quantitative, and gives the wavelength of the light as a bonus. This chapter develops that treatment carefully, and the next takes up the closely related phenomenon of diffraction. Together they put the wave nature of light on the same firm experimental footing that Chapters 1–3 gave to the constancy of its speed — which makes it all the more startling when Chapter 6 shows that light *also* behaves as a stream of particle-like quanta, and that Newton was not simply wrong.

4.1 Waves, Huygens's Principle, and Superposition

Light as an Electromagnetic Wave

Everything in this chapter follows from one picture of what light is, so it is worth stating carefully before using it.

A monochromatic (single-wavelength) light wave traveling in the $+x$ direction consists of an electric field oscillating in a direction perpendicular to x,

$$E(x,t) = E_0 \sin(kx - \omega t + \varphi_0), \tag{4.1}$$

accompanied by a magnetic field perpendicular to both. Here E_0 is the **amplitude**, $k = 2\pi/\lambda$ is the **wave number**, $\omega = 2\pi f$ is the **angular frequency**, and φ_0 is a constant that fixes where in its cycle the wave happens to be at $x = 0$, $t = 0$. The quantity in parentheses is the **phase**. Wavelength, frequency, and speed are linked in the usual way,

$$c = f\lambda. \tag{4.2}$$

Three consequences of this description are used constantly below.

Detectors measure intensity, not field. The **intensity** I — the power per unit area carried by the wave — is proportional to the *time average of the square* of the electric field, and hence to the square of the amplitude:

$$I \propto E_0^2. \tag{4.3}$$

This proportionality is the single most important fact in the chapter. It means amplitudes add, and *then* you square; it is emphatically not true that intensities add. Two waves whose fields cancel produce zero intensity even though each alone carries energy, and (as the phasor section shows) two waves whose fields reinforce produce four times a single wave's intensity, not twice.

> The amplitude E_0 carries units of volts per meter, and the exact relation is $I = \frac{1}{2}c\epsilon_0 E_0^2$ for a wave in vacuum. Only the proportionality $I \propto E_0^2$ is used in this chapter, since every result here compares one intensity to another and the constant $\frac{1}{2}c\epsilon_0$ cancels.

Why the time average? Visible light has $f \approx 5 \times 10^{14}$ Hz, so one oscillation takes about 2×10^{-15} s. No detector — no photodiode, no photographic emulsion, and certainly not the eye — responds anywhere near that fast. Every measurement is an average over an enormous number of cycles, and what survives that average is the intensity.

Light slows down in matter, and its wavelength shrinks with it. In a transparent medium of refractive index n, light travels at speed $v = c/n$. The frequency cannot change — the field at the boundary must oscillate in step on both sides — so the wavelength must:

$$\lambda_n = \frac{\lambda}{n}, \tag{4.4}$$

where λ is the vacuum wavelength. In water ($n = 1.33$), 550 nm green light has a wavelength of only 414 nm.

Optical path length is what counts. Because interference depends on how many *wavelengths* fit into a path, and a medium packs more wavelengths into the same physical distance, the useful quantity is the **optical path length**: for a geometric distance d traversed in a medium of index n,

$$\text{optical path length} = nd. \tag{4.5}$$

A phase advance is 2π per wavelength, so the phase accumulated over a physical distance d in a medium of index n is $(2\pi/\lambda)\, nd$ with λ the *vacuum* wavelength. This idea reappears in the thin-film section and again in the Michelson interferometer, and it is the reason the thin-film formulas below carry a factor of n.

The picture behind all of this — a transverse electric field, a magnetic field perpendicular to it, and both perpendicular to the direction of travel — is drawn in three dimensions and set in motion in Figure 4.1. It is also the place to see what the phase constant φ_0 and the amplitude E_0 actually do, and what it means for a wave to be polarized: a direction of oscillation that plays no part in this chapter, where all the interfering beams share one, but that becomes the whole subject once a filter is put in the beam.

Why Wave Effects Are So Hard to See

Before doing any interference calculation, it pays to understand why nobody notices these effects while walking around. The argument is a scaling argument, and it requires almost no algebra.

Suppose light of wavelength λ passes through an aperture of width a and spreads by some angle θ. What can θ possibly depend on? The problem contains exactly two lengths, λ and a,

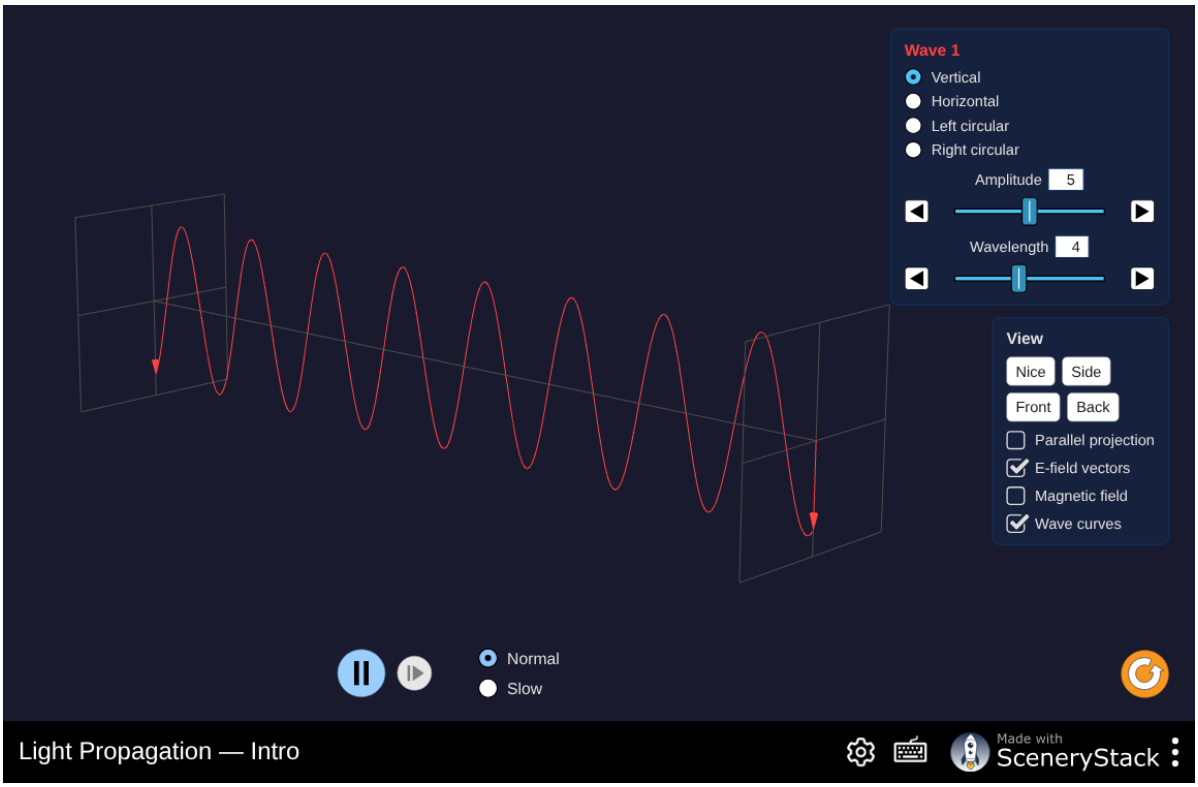

Figure 4.1: A monochromatic electromagnetic wave with its electric and magnetic fields drawn perpendicular to each other and to the direction of propagation. Later screens send the wave through polarizers and birefringent plates, where two components of one wave are given different optical paths — the same mechanism as the thin films of this chapter, applied to polarization rather than to geometry. Interactive simulation: Light Propagation

and θ is a pure number. A pure number cannot depend on a length, so it can only depend on the *ratio* of the two:

$$\theta = (\text{some function of } \lambda/a). \tag{4.6}$$

This is worth pausing on, because it says something strong. Shrink the wavelength and the aperture by the same factor and nothing changes — the whole pattern simply scales. It also means that experiments with water waves in a ripple tank, where wavelengths and apertures are both centimeters, are quantitatively informative about light, where both are tens of thousands of times smaller. What matters is never λ alone; it is λ/a.

The next chapter shows that for a slit the answer is $\theta \approx \lambda/a$, with no large numerical factor. Take that on trust for a moment and put in numbers. For green light, $\lambda = 550$ nm:

- Through a 1 mm pinhole, $\lambda/a = 5.5 \times 10^{-4}$ rad $\approx 0.03°$. Over a 2 m throw the beam spreads by about 1 mm — comparable to the pinhole itself, and easily mistaken for the geometric shadow.
- Through a gap the width of a human hair, $a \approx 70$ μm, the spread is $\lambda/a = 7.9 \times 10^{-3}$ rad, giving about 1.6 cm at 2 m. That is unmistakable, and it is why you can see diffraction by looking at a distant streetlight through a narrow gap between two fingers.
- Through a doorway, $a \approx 1$ m, the spread is 5.5×10^{-7} rad: about 1 μm across a room. Utterly undetectable, which is why doorways cast sharp shadows.

So wave optics does not contradict the ray optics of an introductory course; it *contains* it. Ray optics is what wave optics becomes in the limit $\lambda/a \to 0$, and for everyday objects $\lambda/a \sim 10^{-4}$ or smaller. This is an instance of the **correspondence principle**, an idea that will recur throughout this book: a more general theory must reproduce the older, more limited one in the regime where the older one was tested. Relativity reduces to Newtonian mechanics when $v/c \to 0$ (Chapter 2); quantum mechanics reduces to classical mechanics in the limit of large quantum numbers (Chapter 8); and wave optics reduces to ray optics when $\lambda/a \to 0$.

The practical lesson is simple: to see interference and diffraction, build apertures no more than a few hundred wavelengths across — that is, tenths of a millimeter or smaller. Young's achievement was in large part an achievement of craftsmanship.

Huygens's Principle

Huygens's principle (1678) states that *every point on a wavefront may be treated as a source of secondary spherical wavelets, spreading out at the wave's speed; the wavefront at a later time is the envelope — the surface tangent to all of these wavelets.*

This is a purely geometrical recipe, proposed nearly two centuries before Maxwell's electromagnetic theory existed, and it is remarkably powerful. Applied to a plane wave in open space (Figure 4.2, panel a) it simply regenerates a plane wave, which is the least it must do. Applied at an interface it reproduces the laws of reflection and refraction. And applied at an aperture it makes a prediction that the ray picture cannot: the wavelets emitted near the edges have no neighbors to cancel their sideways spread, so the wave must bend into the geometric shadow. How much it bends depends, as the scaling argument above requires, on λ/a (panels b and c).

Two honest caveats. First, taken literally the construction also predicts a backward-traveling wave, which does not exist; Fresnel and later Kirchhoff repaired this by attaching an *obliquity factor* to each wavelet that suppresses backward emission, and by insisting that the wavelets be added *with their phases* rather than merely enveloped. That upgraded version — the Huygens–Fresnel principle — is the tool of Chapter 5. Second, Huygens's principle by itself

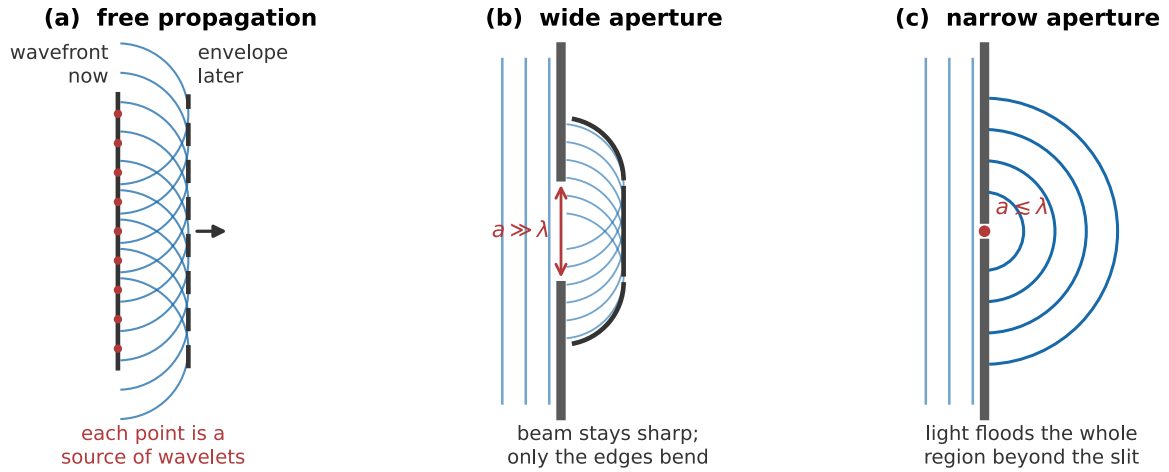

Figure 4.2: Huygens's construction. (a) Each point of a wavefront emits a spherical wavelet; the envelope of the wavelets is the wavefront an instant later. (b) At an aperture much wider than a wavelength, the wavelets in the middle rebuild a flat wavefront and only the edges bend, so the beam looks like a ray. (c) When the aperture is comparable to a wavelength, a single wavelet survives and the transmitted wave spreads over the whole region beyond the slit. Original schematic generated with matplotlib; see `scripts/figures/`.

says nothing about intensity. To get intensities we need the superposition principle, which is where we turn next.

Superposition: Turning Path Difference into Phase Difference

The **superposition principle** states that when two or more waves overlap, the resulting disturbance at each point and each instant is the *sum* of the individual disturbances. For light this is a consequence of the linearity of Maxwell's equations in vacuum and in ordinary transparent materials: fields add, and they add as vectors.

Consider two waves of the same frequency and the same amplitude E_0 arriving at some point P. Suppose they left a common source in step with each other but then traveled different distances, r_1 and r_2, to get to P. Each accumulates phase at a rate of 2π radians per wavelength traveled, so their **phase difference** at P is

$$\phi = \frac{2\pi}{\lambda}\,(r_2 - r_1) = \frac{2\pi}{\lambda}\,\Delta r. \tag{4.7}$$

This one equation is the hinge of the whole chapter, and it is worth reading in both directions. A path difference of one whole wavelength corresponds to a phase difference of 2π, which is no phase difference at all — the waves are back in step. A path difference of half a wavelength corresponds to a phase difference of π, which puts crest against trough. In general:

$$\Delta r = m\lambda \quad (m = 0, \pm 1, \pm 2, \ldots) \qquad \Longrightarrow \qquad \phi = 2\pi m \qquad \textbf{constructive}, \tag{4.8}$$

$$\Delta r = \left(m + \tfrac{1}{2}\right)\lambda \qquad \Longrightarrow \qquad \phi = (2m+1)\pi \qquad \textbf{destructive}. \tag{4.9}$$

Under constructive interference the fields add crest to crest, giving a resultant amplitude $2E_0$ and — since $I \propto E_0^2$ — an intensity *four* times that of either wave alone. Under destructive

interference the resultant amplitude is $E_0 - E_0 = 0$ and the intensity vanishes. If the two amplitudes are unequal, the extremes are $E_1 + E_2$ and $|E_1 - E_2|$; complete darkness requires equal amplitudes, which is why interference fringes look best when the two slits are identical.

Three sources of phase difference will appear in this chapter, and it is worth naming them now so that they are not confused later:

1. **Path difference** in vacuum or air, $\phi = (2\pi/\lambda)\Delta r$ — the double slit.
2. **Optical path difference** when part of the path lies in a medium, $\phi = (2\pi/\lambda)\,\Delta(nd)$ — thin films, and gas cells in an interferometer.
3. **Phase shifts on reflection**, an abrupt π acquired at certain boundaries — thin films again.

Every interference problem in this chapter is solved by totaling these three contributions and asking whether the total is an even or an odd multiple of π.

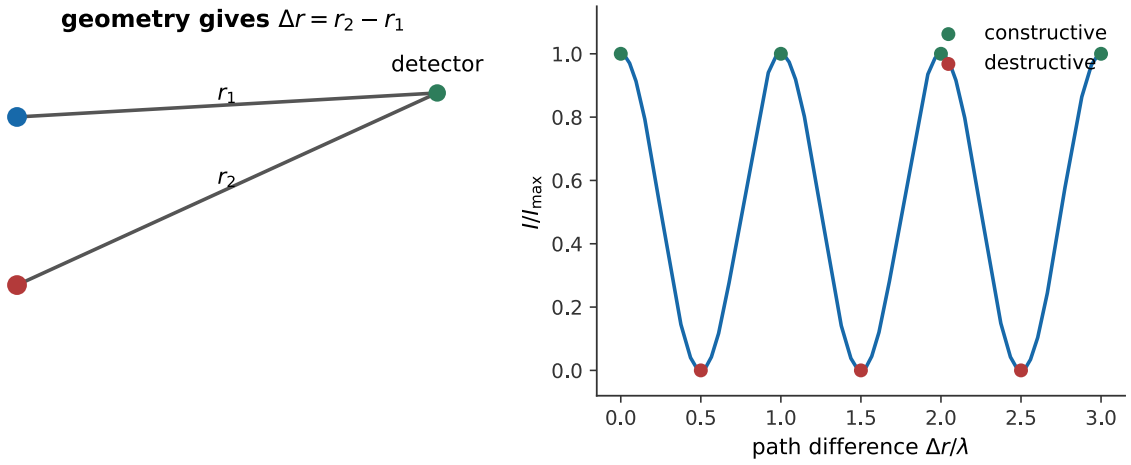

Figure 4.3: Move the detector across the field and watch geometry turn continuously into phase. Each whole-wavelength change in path difference returns the waves to constructive interference; each half-wavelength puts them in opposition. The same sweep becomes the bright and dark bands on Young's screen.

Worked Example: From Path Difference to Brightness

Two identical, in-step sources emit light of wavelength $\lambda = 500$ nm. At a particular point the waves have traveled distances differing by $\Delta r = 1.25$ μm. Is that point bright or dark?

Count wavelengths in the path difference:

$$\frac{\Delta r}{\lambda} = \frac{1.25 \times 10^{-6}\ \text{m}}{500 \times 10^{-9}\ \text{m}} = 2.5. \tag{4.10}$$

The path difference is two and a half wavelengths — a half-integer number — so $\phi = 2\pi(2.5) = 5\pi$, an odd multiple of π. The point is **dark**.

Now suppose the whole region between the sources and that point is filled with water, $n = 1.33$, without moving anything. The geometric path difference is unchanged, but the wavelength in the water is $\lambda_n = (500\ \text{nm})/1.33 = 376$ nm, so the path difference is now

(1250 nm)/(376 nm) = 3.32 wavelengths and the point is neither fully bright nor fully dark. Immersing an interference experiment in a medium genuinely changes the pattern; it does not merely rescale the brightness.

4.2 Young's Double Slit and Coherence

Young's Double-Slit Experiment

Young's arrangement is shown in Figure 4.4. Light from a source passes first through a *single* narrow slit, then through *two* narrow slits separated by a distance d, and finally falls on a screen a distance L away.

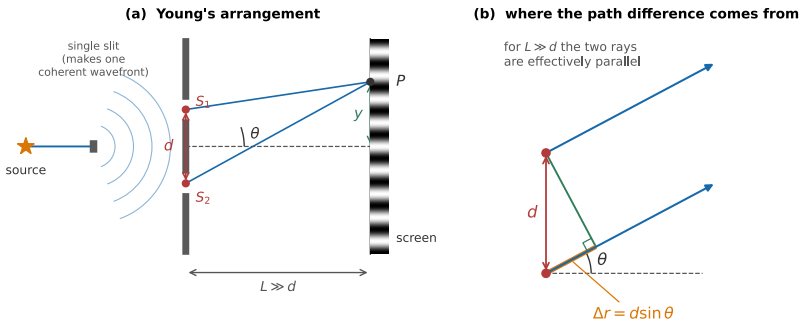

Figure 4.4: Young's double-slit experiment. (a) The first slit produces a single wavefront that illuminates both of the second pair, so S_1 and S_2 act as coherent sources. (b) For $L \gg d$ the two rays reaching a distant point are effectively parallel, and the ray from S_2 travels an extra distance $\Delta r = d \sin\theta$. Original schematic generated with matplotlib; see `scripts/figures/`.

The first slit is not decoration. Its job is to guarantee that a single wavefront reaches both of the following slits, so that whatever the source does — however erratically it flickers — it does the same thing at S_1 and at S_2 at the same moment. The two slits then behave as **coherent** sources with a fixed phase relationship. Without that first slit, different parts of an extended source illuminate the two slits independently, and the pattern washes out. (The coherence section below makes this precise; a modern laser has enough built-in coherence that the first slit can be omitted, which is why the classroom demonstration looks so much easier than Young's original.)

> **Note**
> Young's actual 1801–1803 experiments looked little like the tidy two-slit diagram of Figure 4.4. He first produced two coherent beams by letting sunlight through a single pinhole fall on the two edges of a thin card held edge-on in the beam, and only later described the equivalent double-slit arrangement in print. Manufacturing two slits narrow and close enough to show clean fringes was, at the time, a serious feat of craftsmanship; Augustin-Jean Fresnel, extending the work a decade later, devised alternative ways to split a single wavefront — his double mirror, and his biprism, a single piece of glass ground into two shallow prisms back to back — that avoided a fragile double slit altogether while producing exactly the same fringe pattern. The physics is identical in all three arrangements:



Figure 4.5: A recreation of Young's original apparatus, framed around the same wave-versus-particle question this chapter answers — and that Chapter 7 reopens for matter itself. Watch on YouTube: The Original Double Slit Experiment

what matters is only that two coherent virtual sources end up separated by a controllable distance d.

The Path Difference

Consider a point P on the screen, at angle θ from the central axis. Light reaching P from S_2 has traveled farther than light from S_1. Because $L \gg d$ in any practical apparatus — d is tens of micrometers, L is a meter or more — the two rays heading for P are very nearly parallel, and the geometry collapses to the small right triangle of Figure 4.4: drop a perpendicular from S_1 onto the ray from S_2, and the extra leg is

$$\Delta r = d \sin\theta. \tag{4.11}$$

Combining this with the constructive and destructive conditions of the previous section gives the **two-slit interference conditions**:

$$d \sin\theta = m\lambda \qquad \text{(bright fringe)}, \qquad m = 0, \pm 1, \pm 2, \ldots \tag{4.12}$$

$$d \sin\theta = \left(m + \tfrac{1}{2}\right)\lambda \qquad \text{(dark fringe)}. \tag{4.13}$$

The integer m is the **order** of the fringe. The $m = 0$ bright fringe sits on the axis, where the two paths are exactly equal; note that its position does not depend on λ, so in white light the central fringe is white while all the others are spread into little spectra.

Fringe Positions on the Screen

For small angles — again, the usual case — $\sin\theta \approx \tan\theta \approx y/L$, where y is measured on the screen from the central axis. The bright fringes then fall at

$$y_m = \frac{m\lambda L}{d}, \tag{4.14}$$

evenly spaced, with **fringe spacing**

$$\Delta y = \frac{\lambda L}{d}. \tag{4.15}$$

Three features of this result deserve comment. First, measuring Δy, L, and d determines λ — this was historically one of the first accurate measurements of the wavelength of visible light, and the first evidence that different colors correspond to different, specific wavelengths. Second, the fringes get *farther apart* as the slits get *closer together*: squeezing the apparatus stretches the pattern. This reciprocal relationship between a wave's confinement and its spread is a hallmark of wave phenomena, and it returns as the uncertainty principle in Chapter 7. Third, the small-angle form is an approximation; when d is only a few wavelengths the fringes are not evenly spaced and $d \sin\theta = m\lambda$ must be used directly.

Tip

Before using $\Delta y = \lambda L/d$, check the approximation numerically rather than by eye: compute d/L and confirm it is small (a few times 10^{-4} or smaller is typical for a tabletop double slit), and check that the order m of interest keeps $\sin\theta = m\lambda/d$ well under 1. Both conditions hold comfortably in every worked example below, but a request for a high-order fringe, or an unusually large d/L, can push $\sin\theta$ far enough from θ that only the exact condition $d\sin\theta = m\lambda$ remains trustworthy.

The reciprocal relationship in $\Delta y = \lambda L/d$ is easy to state and easy to get backwards, so it is worth watching it happen. In Figure 4.6 the slit separation is a slider and the pattern responds live: closing the slits together spreads the fringes apart, and raising the frequency — shortening λ — packs them closer. The simulation runs the same geometry with water waves, sound, and light, which is a useful reminder that nothing in the derivation used any property of light beyond its being a wave with a wavelength.

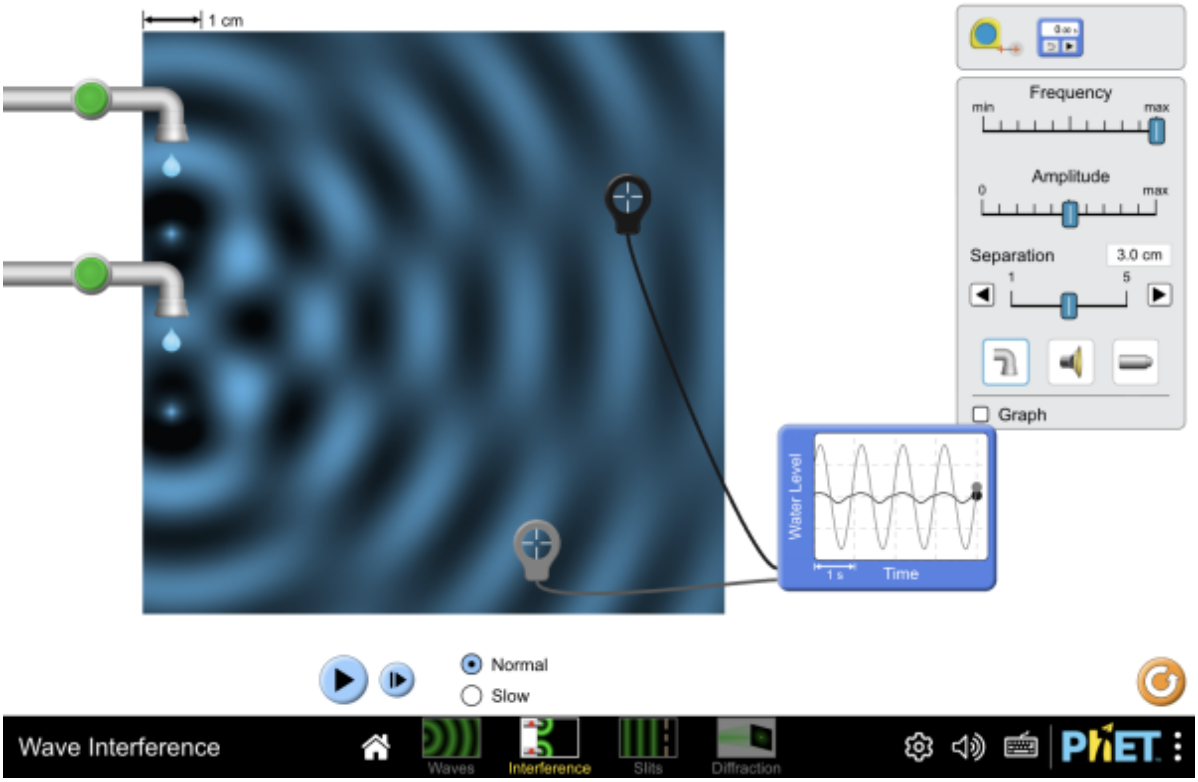

Figure 4.6: Young's geometry with the slit separation, the slit width, and the wavelength under direct control. Switch between one slit and two: with one slit open the screen still shows structure, which is the diffraction of Chapter 5 intruding on the idealization used here, where the slits are treated as point sources. Interactive simulation: Wave Interference

Finally, since $\sin\theta$ can never exceed 1, the condition $d\sin\theta = m\lambda$ has solutions only for

$$|m| \le \frac{d}{\lambda}, \tag{4.16}$$

so a double slit produces a finite number of orders — a point that becomes important for diffraction gratings in Chapter 5.

Worked Example: Wavelength from Fringe Spacing

Slits separated by $d = 0.200$ mm are illuminated by a laser, and the resulting fringes on a screen $L = 2.00$ m away are found to be spaced $\Delta y = 6.50$ mm apart. Find the laser wavelength.

Solving $\Delta y = \lambda L/d$ for λ:

$$\lambda = \frac{\Delta y\, d}{L} = \frac{(6.50 \times 10^{-3}\ \text{m})(0.200 \times 10^{-3}\ \text{m})}{2.00\ \text{m}} = 6.50 \times 10^{-7}\ \text{m} = 650\ \text{nm}, \tag{4.17}$$

consistent with a red helium–neon or diode laser. Note the sizes of the quantities involved: a 650 nm wavelength has been measured with a millimeter ruler, because the apparatus magnifies the wavelength by the factor $L/d = 10^4$.

Worked Example: Fringe Angles, Positions, and Count

A double slit with $d = 0.100$ mm is illuminated with the green line of a mercury lamp, $\lambda = 546$ nm, and the screen is $L = 1.20$ m away. (a) Find the angle to the third-order bright fringe. (b) Find the fringe spacing on the screen. (c) How many bright fringes appear on a screen 5.0 cm wide, centered on the axis? (d) How many orders exist in principle?

(a) From $d \sin\theta = m\lambda$ with $m = 3$:

$$\sin\theta_3 = \frac{3(546 \times 10^{-9}\ \text{m})}{1.00 \times 10^{-4}\ \text{m}} = 0.0164 \quad \Longrightarrow \quad \theta_3 = 0.939^\circ. \tag{4.18}$$

The angle is small, which retroactively justifies the small-angle work in (b).

(b)

$$\Delta y = \frac{\lambda L}{d} = \frac{(546 \times 10^{-9}\ \text{m})(1.20\ \text{m})}{1.00 \times 10^{-4}\ \text{m}} = 6.55 \times 10^{-3}\ \text{m} = 6.55\ \text{mm}. \tag{4.19}$$

(c) The screen extends from $y = -2.5$ cm to $+2.5$ cm, so the highest order that lands on it satisfies $|m|\Delta y \le 25$ mm, giving $|m| \le (25\ \text{mm})/(6.55\ \text{mm}) = 3.8$. Orders $m = -3$ through $m = +3$ appear: **seven bright fringes**.

(d) In principle, $|m| \le d/\lambda = (1.00 \times 10^{-4}\ \text{m})/(546 \times 10^{-9}\ \text{m}) = 183$. The far orders are useless in practice — they lie at large angles where the small-angle formula fails, and (as Chapter 5 shows) the finite width of real slits has long since dimmed them to nothing — but the counting matters for gratings.

Worked Example: The Same Experiment Under Water

The apparatus of the previous example is submerged in water, $n = 1.33$, source and screen included. What happens to the fringe spacing?

The frequency of the light does not change, but its wavelength does: $\lambda_n = \lambda/n = (546\ \text{nm})/1.33 = 411$ nm. Nothing in the derivation of $\Delta y = \lambda L/d$ referred to vacuum, so the same formula holds with the *local* wavelength:

$$\Delta y_{\text{water}} = \frac{\lambda_n L}{d} = \frac{\Delta y}{n} = \frac{6.55\ \text{mm}}{1.33} = 4.93\ \text{mm}. \tag{4.20}$$

The pattern contracts by the factor n. This is a genuinely useful check on understanding: it is the wavelength *where the interference happens* that sets the scale, and that is also why the thin-film formulas below carry a factor of n_{film}.

Coherence

Young's experiment works only because both slits are carved out of the *same* wavefront, so the light leaving them keeps a fixed phase relationship. Try the experiment with two separate light bulbs and you see nothing but a uniformly lit screen. The reason is not that light bulbs are dim; it is that they are **incoherent**.

An ordinary thermal source — a filament, a flame, a fluorescent tube — emits light as an enormous number of independent atomic events. Each excited atom radiates a short burst, a **wave train** lasting perhaps 10^{-8} s, and the next atom to radiate does so with no memory of the phase of the last. What emerges is therefore a chain of wave trains whose phase jumps randomly every few nanoseconds, as sketched in Figure 4.7.

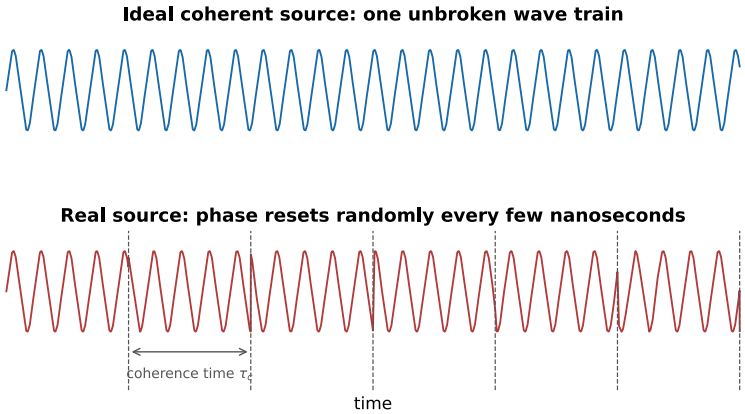

Figure 4.7: Coherence. An ideal source (top) maintains a predictable phase indefinitely. A real thermal source (bottom) emits short wave trains whose phase resets at random every coherence time τ_c; any interference pattern formed by one train is replaced by a differently positioned pattern from the next, far faster than a detector can follow. Original schematic generated with matplotlib; see `scripts/figures/`.

Two independent sources therefore do produce an interference pattern — but a different one every few nanoseconds, in a random new position each time. Averaged over the nanoseconds to milliseconds that any real detector needs, the fringes wash out completely and the intensities simply add. The pattern is not weak; it is scrambled.

Temporal Coherence, Coherence Time, and Coherence Length

Two quantities measure how long a source stays predictable:

- The **coherence time** τ_c: roughly the duration of one uninterrupted wave train, the time over which the phase remains predictable.
- The **coherence length** $\ell_c = c\,\tau_c$: the corresponding path-length difference over which fringes survive. If the two arms of an interference experiment differ by much more than ℓ_c, the light recombining at the screen comes from two *different*, unrelated wave trains, and there is nothing stable to interfere.

> A source is called **quasi-monochromatic** when its spectral width satisfies $\Delta\lambda \ll \lambda$ — true, to greater or lesser degree, of every source in this section except the explicitly white-light

Table 4.1: Representative coherence lengths for common light sources

Source	λ	$\Delta\lambda$	$\ell_c \approx \lambda^2/\Delta\lambda$
White light	550 nm	~ 300 nm	~ 1 μm (about two wavelengths)
Filtered sodium lamp	589 nm	~ 0.6 nm	~ 0.6 mm
Helium–neon laser	633 nm	~ 0.002 nm	~ 20 cm
Stabilized single-mode laser	633 nm	$\sim 10^{-6}$ nm	hundreds of meters

examples. It is this condition that makes "the wavelength," singular, a meaningful thing to quote at all.

Coherence length is set by the spectral purity of the source. A wave train of finite duration τ_c cannot be perfectly monochromatic — Fourier analysis gives it a spread of frequencies $\Delta f \sim 1/\tau_c$ — and converting to wavelength gives the useful estimate

$$\ell_c \approx \frac{\lambda^2}{\Delta\lambda}. \tag{4.21}$$

The representative values in Table 4.1 span an extraordinary range:

A closely related way to say the same thing: the number of fringes you can count before the pattern fades is roughly

$$N_{\text{fringes}} \approx \frac{\ell_c}{\lambda} \approx \frac{\lambda}{\Delta\lambda}. \tag{4.22}$$

White light gives you two or three fringes — which is exactly what Young saw, and exactly why the colored fringes of a soap bubble are confined to a film only a few wavelengths thick. A helium–neon laser gives several hundred thousand.

Spatial Coherence

Temporal coherence is about a single point in the beam staying predictable *over time*. **Spatial coherence** is about two *different* points across the beam having a fixed phase relationship *at the same time*, and it is what Young's first slit provides. An extended source — the Sun's disk, a broad filament — has poor spatial coherence, because light arriving at S_1 and light arriving at S_2 come from different, unrelated emitters. Passing the light through a pinhole first discards almost all of it but leaves what remains spatially coherent, since it all now originates from a region small enough to act as a single point.

A laser is coherent in both senses at once, which is why a laser pointer and a pair of razor-blade slits reproduce in seconds an experiment that took Young considerable ingenuity.

Worked Example: Will the Fringes Survive?

A Michelson interferometer (below) has arms differing in length by $\Delta L = 5.0$ cm, so the two recombining beams differ in path by $2\Delta L = 10$ cm. Will fringes be visible with (a) a filtered sodium lamp, (b) a helium–neon laser?

(a) From Table 4.1, $\ell_c \approx 0.6$ mm for the sodium lamp. The path difference of 10 cm exceeds this by a factor of about 170, so the beams arriving together come from unrelated wave trains: **no fringes**.

(b) For the laser, $\ell_c \approx 20$ cm, comfortably larger than the 10 cm path difference: **fringes are visible**, though with reduced contrast since the path difference is a sizable fraction of ℓ_c.

This is not a contrived exercise. Michelson had to keep his arms equal to within a fraction of a millimeter precisely because his sodium light had a coherence length of well under a millimeter — a serious experimental constraint in 1887, and one that vanished with the invention of the laser.

4.3 Intensity and Multiple Slits

Intensity: Adding Fields as Phasors

The conditions derived so far locate the bright and dark fringes but say nothing about the brightness in between. To get the full pattern we must add the two waves properly, and the most convenient bookkeeping for that is the **phasor**.

Because both waves have the same frequency, their relative phase does not change with time; only their common phase does. Represent each wave by a vector (a phasor) whose length is the wave's amplitude and whose direction gives its phase. The whole diagram rotates rigidly at the optical frequency, so we may freeze it at any instant; adding the waves is then just adding the vectors, and the length of the resultant vector is the amplitude of the combined wave.

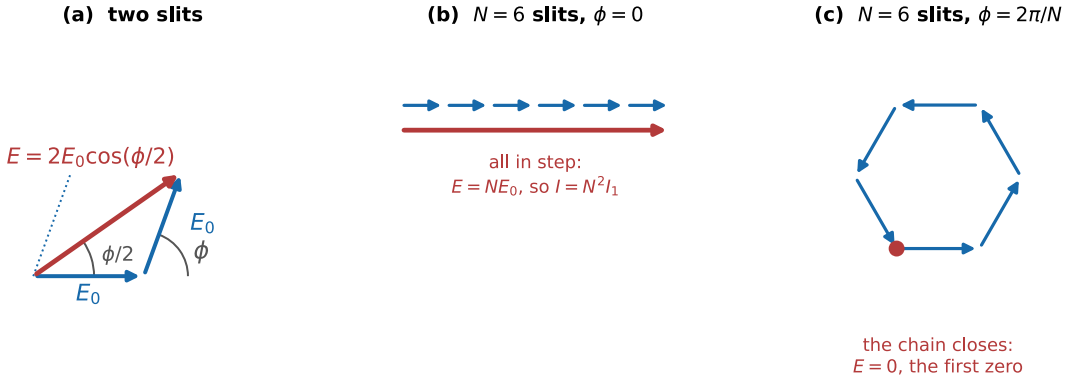

Figure 4.8: Phasor addition. (a) Two equal phasors separated by ϕ form an isosceles triangle whose base is the resultant, $E = 2E_0\cos(\phi/2)$. (b) With N slits and $\phi = 0$ all the phasors line up, giving $E = NE_0$ and hence $I = N^2 I_1$. (c) When $\phi = 2\pi/N$ the chain closes on itself and the resultant vanishes — the first zero, at only $1/N$ of the way to the next principal maximum. Original schematic generated with matplotlib; see `scripts/figures/`.

Two Slits

For the double slit, two phasors of equal length E_0 are separated by the angle

$$\phi = \frac{2\pi}{\lambda}\, d \sin\theta. \tag{4.23}$$

They form an isosceles triangle (Figure 4.8). The resultant bisects the angle between them, and dropping a perpendicular gives immediately

$$E = 2E_0 \cos\left(\frac{\phi}{2}\right). \tag{4.24}$$

Squaring, and writing $I_1 \propto E_0^2$ for the intensity one slit alone would produce,

$$I(\theta) = 4I_1 \cos^2\left(\frac{\phi}{2}\right) = I_0 \cos^2\left(\frac{\pi d \sin\theta}{\lambda}\right), \qquad I_0 \equiv 4I_1. \tag{4.25}$$

The result is plotted in Figure 4.9. It reproduces everything derived earlier — $I = I_0$ whenever $d\sin\theta = m\lambda$, and $I = 0$ whenever $d\sin\theta = (m + \frac{1}{2})\lambda$ — and now fills in the smooth $\cos^2$ variation between.

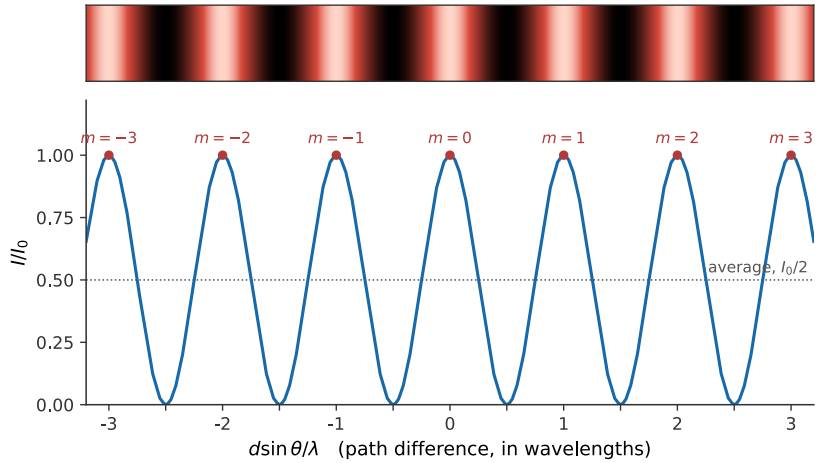

Figure 4.9: Two-slit intensity, $I = I_0 \cos^2(\pi d \sin\theta/\lambda)$, with the corresponding fringe pattern above. All the maxima have the same height and the fringes are evenly spaced — a signature of ideal, infinitesimally narrow slits. Real slits impose the envelope derived in Chapter 5. Generated with matplotlib; see `scripts/figures/`.

Where Does the Extra Light Come From?

The central maximum has intensity $I_0 = 4I_1$: four times what *one* slit would deliver, not twice. Students often find this alarming, and they should — until they check the energy budget.

The resolution is that $\cos^2$ averages to $\frac{1}{2}$. Averaged across the pattern, the intensity is

$$\langle I \rangle = 4I_1 \times \tfrac{1}{2} = 2I_1, \tag{4.26}$$

which is exactly the two slits' worth of light that entered. Interference does not create or destroy energy; it **redistributes** it, taking light from the dark fringes and piling it into the bright ones. The factor of four at the peaks is paid for by the zeros in between.

This is worth stating sharply, because it is the cleanest possible refutation of a classical particle picture of light: opening a second slit makes some places on the screen *darker* than they were with one slit open. No stream of independent corpuscles can do that.

Worked Example: Intensity Between the Fringes

For the mercury-lamp double slit above ($d = 0.100$ mm, $\lambda = 546$ nm, $L = 1.20$ m), at what distance from the center of the pattern does the intensity fall to half its maximum?

Set $I = I_0/2$:

$$\cos^2\left(\frac{\pi d \sin\theta}{\lambda}\right) = \frac{1}{2} \quad \Longrightarrow \quad \frac{\pi d \sin\theta}{\lambda} = \frac{\pi}{4} \quad \Longrightarrow \quad d\sin\theta = \frac{\lambda}{4}. \tag{4.27}$$

So the half-intensity point occurs at a quarter-wavelength path difference — halfway from a bright fringe to the next dark one. In small-angle terms $y = \Delta y/4$:

$$y = \frac{6.55 \text{ mm}}{4} = 1.64 \text{ mm}. \tag{4.28}$$

The full width at half maximum of each fringe is therefore $2y = 3.28$ mm, exactly half the fringe spacing in this small-angle approximation. Two-slit fringes are broad and sinusoidal — which, as the next section shows, is precisely what more slits fix.

Three Slits, and Then N

Add a third identical slit, equally spaced. What changes?

The principal maxima do not move. At $d\sin\theta = m\lambda$, *every* slit is in step with every other, so all three phasors point the same way, the resultant is $3E_0$, and the intensity is $9I_1$. Adding slits at the same spacing can never shift these directions; it just adds more phasors to an already aligned stack.

What changes is where the *first zero* falls. With two slits, the resultant vanishes when the second phasor is antiparallel to the first: $\phi = \pi$. With three slits, the three phasors close into an equilateral triangle — and hence sum to zero — as soon as $\phi = 2\pi/3$, well before ϕ reaches π. The maximum has become narrower. With N slits, the chain closes into a regular N-gon at $\phi = 2\pi/N$ (Figure 4.8), so the first zero sits only $1/N$ of the way to the next principal maximum.

Summing the general chain is a geometric series. Writing each successive phasor as the previous one multiplied by $e^{i\phi}$, the total field is $E_0(1+e^{i\phi}+e^{2i\phi}+\cdots+e^{i(N-1)\phi})$, whose magnitude works out to $E_0\,|\sin(N\phi/2)/\sin(\phi/2)|$. Squaring gives the **$N$-slit intensity pattern**

$$I(\theta) = I_1\left[\frac{\sin(N\phi/2)}{\sin(\phi/2)}\right]^2, \qquad \phi = \frac{2\pi}{\lambda} d\sin\theta, \tag{4.29}$$

where I_1 is the intensity from a single slit. Setting $N = 2$ recovers the two-slit result, since $\sin\phi/\sin(\phi/2) = 2\cos(\phi/2)$.

Summing the N-slit phasor chain

The total field at angle θ is the sum of N equal-amplitude phasors, each rotated by ϕ from the last:

$$E = E_0\left(1 + e^{i\phi} + e^{2i\phi} + \cdots + e^{i(N-1)\phi}\right) = E_0\sum_{k=0}^{N-1} e^{ik\phi}. \tag{4.30}$$

This is a finite geometric series with common ratio $e^{i\phi}$, so

$$E = E_0\,\frac{1 - e^{iN\phi}}{1 - e^{i\phi}} \qquad (\phi \neq 0). \tag{4.31}$$

To get the magnitude, factor a half-angle phase out of both numerator and denominator: for any angle α, $1 - e^{i\alpha} = e^{i\alpha/2}\left(e^{-i\alpha/2} - e^{i\alpha/2}\right) = -2i\,e^{i\alpha/2}\sin(\alpha/2)$. Applying this with $\alpha = N\phi$ in the numerator and $\alpha = \phi$ in the denominator,

$$E = E_0\,\frac{-2i\,e^{iN\phi/2}\sin(N\phi/2)}{-2i\,e^{i\phi/2}\sin(\phi/2)} = E_0\,e^{i(N-1)\phi/2}\,\frac{\sin(N\phi/2)}{\sin(\phi/2)}. \tag{4.32}$$

The prefactor $e^{i(N-1)\phi/2}$ has magnitude 1 — it is only an overall phase, reflecting the arbitrary choice of the first slit as the phase origin — so the amplitude is

$$|E| = E_0\left|\frac{\sin(N\phi/2)}{\sin(\phi/2)}\right|, \tag{4.33}$$

exactly the result quoted in the main text. Squaring gives the N-slit intensity formula, and setting $N = 2$ recovers the two-slit result via the identity $\sin\phi = 2\sin(\phi/2)\cos(\phi/2)$.

Figure 4.10 collects the consequences:

- **Principal maxima** occur wherever ϕ is a multiple of 2π, that is $d\sin\theta = m\lambda$ — the same directions as for two slits, for any N. Their height is $N^2 I_1$.
- **Zeros** occur at $\phi = 2\pi p/N$ for every integer p that is *not* a multiple of N. There are $N-1$ zeros between consecutive principal maxima, and hence $N-2$ much weaker **secondary maxima** squeezed between them.
- **Width.** Since the first zero is at $\phi = 2\pi/N$, the angular half-width of a principal maximum scales as $1/N$: doubling the number of illuminated slits halves the width of every line.

The two scalings work together. The peak height grows as N^2 while the width shrinks as $1/N$, so the energy in each principal maximum grows only as N — as it must, since N slits admit N times as much light. What you gain by adding slits is not more light in total but light concentrated into sharper and sharper spikes at precisely determined angles. Since those angles depend on λ, a device with very many slits becomes an extremely precise wavelength meter. That device is the **diffraction grating**, and it is taken up — together with the complication that real slits have finite width — in Chapter 5.

Worked Example: Locating the Zeros of a Five-Slit Pattern

Five slits with $d = 2.00\ \mu$m are illuminated at $\lambda = 500$ nm. Find the directions of the principal maxima, and of the zeros lying between the central maximum and the first-order maximum.

Principal maxima: $\sin\theta = m\lambda/d = m(500\text{ nm})/(2000\text{ nm}) = 0.250\,m$, giving $m = 0, \pm1, \pm2, \pm3$ at $\sin\theta = 0, 0.250, 0.500, 0.750$, plus $m = \pm4$ exactly at $\sin\theta = 1$ (grazing, and not observable).

Zeros: $\phi = 2\pi p/5$ means $d\sin\theta = p\lambda/5$, so

$$\sin\theta = \frac{p\lambda}{Nd} = \frac{p(500\text{ nm})}{5(2000\text{ nm})} = 0.0500\,p, \tag{4.34}$$

for $p = 1, 2, 3, 4$ (the value $p = 5$ is excluded — it is the first-order principal maximum). The four zeros lie at $\sin\theta = 0.050, 0.100, 0.150, 0.200$, and between them sit $N-2 = 3$ secondary maxima. Note that the first zero is at $\sin\theta = 0.050$, one fifth of the way to the first principal maximum at 0.250. For comparison, the first zero of the two-slit pattern is halfway to that maximum, so this five-slit central peak is 2.5 times narrower than the corresponding two-slit peak.

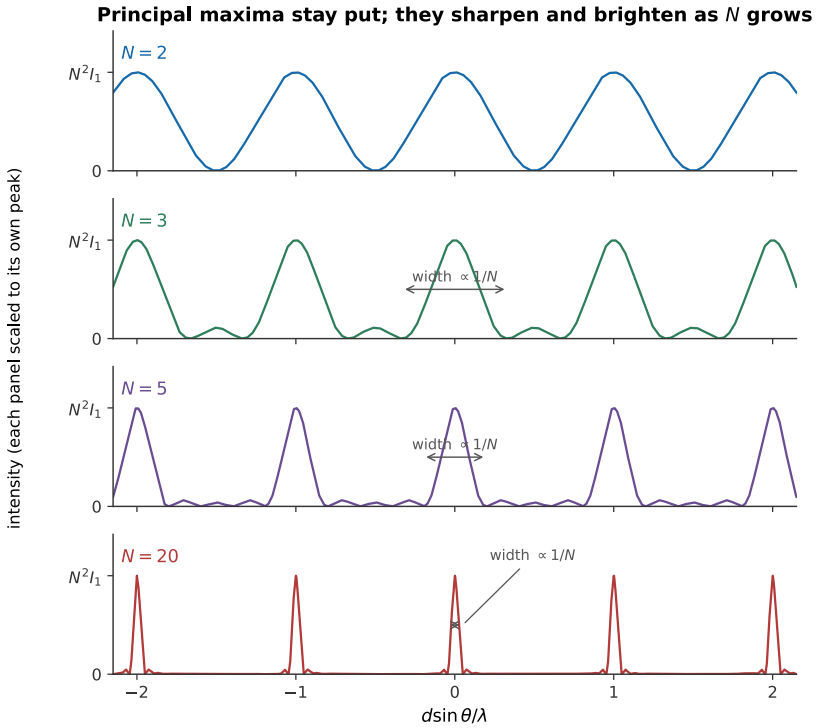

Figure 4.10: Interference from N equally spaced ideal slits. The principal maxima stay at $d\sin\theta = m\lambda$ regardless of N, but they grow taller (peak $\propto N^2$) and narrower (width $\propto 1/N$), and $N-2$ weak secondary maxima appear between them. Each panel is scaled to its own peak. Generated with matplotlib; see `scripts/figures/`.

4.4 Thin Films and Interferometers

Thin-Film Interference

The most familiar interference in everyday life needs no slits at all. The colors swirling on a soap bubble, the rainbow sheen of oil on a wet road, and the faint purple cast of a coated camera lens are all produced by light reflecting from the two surfaces of a very thin transparent layer.

Figure 4.11 shows the situation. Light striking a film of thickness t and refractive index n_{film} partially reflects at the front surface (ray 1) and partially enters the film, reflects from the back surface, and re-emerges (ray 2). The two emerging beams are parallel, and the eye or a lens brings them together to interfere.

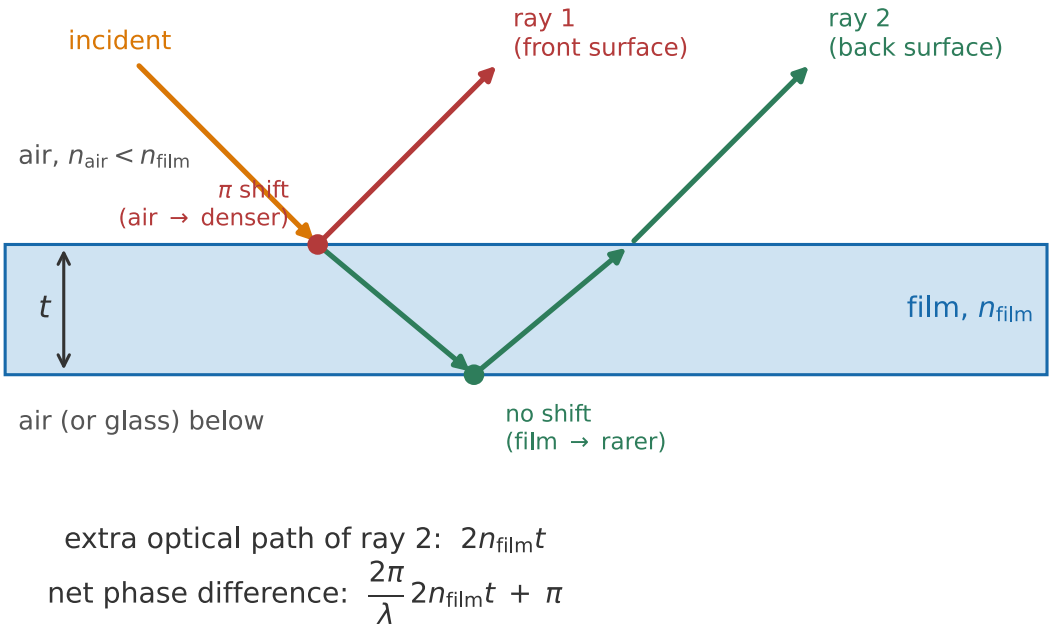

Figure 4.11: Thin-film interference. Ray 1 reflects from the front surface, where the index increases, and picks up a π phase shift. Ray 2 makes a round trip inside the film, acquiring an extra optical path $2n_{\text{film}}t$, and reflects from the back surface, where the index decreases, with no shift. The net phase difference is the sum of the two effects. Original schematic generated with matplotlib; see `scripts/figures/`.

The Two Contributions

Optical path. At near-normal incidence, ray 2 travels an extra geometric distance $2t$, all of it inside the film. By the optical-path rule, this corresponds to a phase difference of

$$\phi_{\text{path}} = \frac{2\pi}{\lambda}\,2n_{\text{film}}t, \tag{4.35}$$

with λ the vacuum wavelength. The factor n_{film} is not optional: the film packs more wavelengths into the same thickness.

Phase shift on reflection. Reflection at a boundary where the refractive index *increases* (light in a lower-index medium striking a higher-index one) flips the sign of the reflected field — an extra phase shift of exactly π. Reflection at a boundary where the index *decreases* produces no shift. The mechanical analog is a wave on a string: a pulse reflecting from a fixed end (a heavy string beyond, the "denser" case) comes back inverted, while a pulse reflecting from a free end (a light string beyond) comes back upright. "Higher refractive index" plays the role of "heavier string".

That analog is worth more than a sentence, because the sign is the one thing students reliably get wrong and it is not something to be memorized. Figure 4.12 launches a pulse down a chain of masses and springs terminated either rigidly or freely, and the inversion is not scripted into the simulation: it falls out of the boundary condition, exactly as the π shift falls out of matching the electric field across an optical interface. A fixed end cannot move, so the reflected pulse must arrive with the opposite sign to cancel the incident one there; a free end has nothing to push against, and the pulse returns upright.

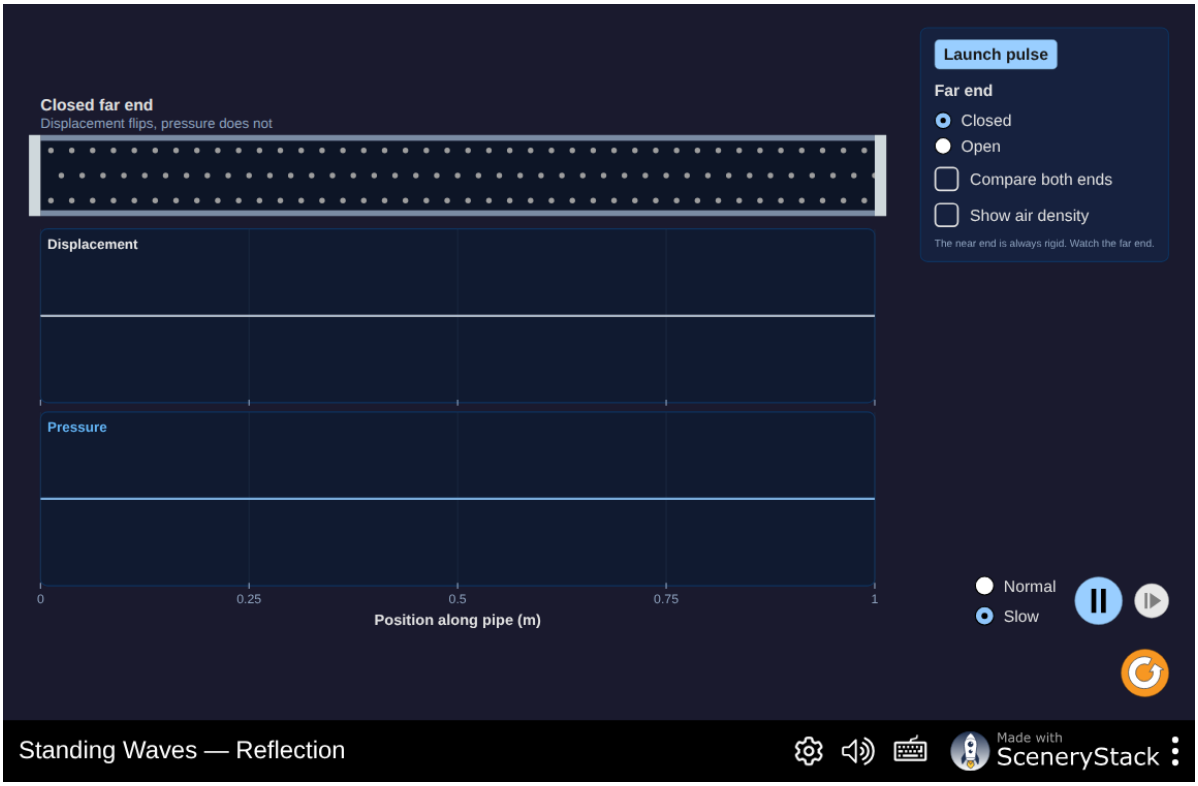

Figure 4.12: A pulse reflecting from a rigid termination and from a free one, side by side on the same clock. The rigid end inverts the pulse and the free end does not — the mechanical statement of the 0-or-π rule used throughout this section. Interactive simulation: Standing Waves

Assembling the Conditions

Because the reflection shifts are either 0 or π, the two reflections in Figure 4.11 can only produce a *net* relative shift of 0 or π. This gives a reliable three-step recipe:

1. Check the front reflection. Does the index increase going into the film? If so, that ray gets π.

2. Check the back reflection. Does the index increase going out of the film into whatever lies beyond? If so, that ray gets π.

3. If exactly one of the two picked up π, the net reflection shift is π; if both did or neither did, it is 0.

Case A — net shift of π (exactly one reflection flips). This covers a soap film in air, an oil slick on water, and an air gap trapped between two glass plates. Bright reflection needs the path contribution to supply the missing half wavelength:

$$2n_{\text{film}}t = \left(m + \tfrac{1}{2}\right)\lambda \qquad \text{(constructive reflection)}, \tag{4.36}$$

$$2n_{\text{film}}t = m\lambda \qquad \text{(destructive reflection)}. \tag{4.37}$$

Case B — net shift of 0 (both flip, or neither does). This covers any film whose index lies between those of the media on either side — an antireflection coating on glass, or an oil film on a denser substrate. The conditions are exactly reversed:

$$2n_{\text{film}}t = m\lambda \qquad \text{(constructive reflection)}, \tag{4.38}$$

$$2n_{\text{film}}t = \left(m + \tfrac{1}{2}\right)\lambda \qquad \text{(destructive reflection)}. \tag{4.39}$$

Getting the case wrong swaps bright for dark everywhere, so it is worth the ten seconds it takes to check.

Why Soap Films Are Colored — and Why They Go Black

Since the conditions involve λ, a film illuminated with white light reflects some wavelengths strongly and suppresses others, and the favored wavelength shifts as the thickness changes. Figure 4.13 shows the reflected intensity of red, green, and blue light as a function of the thickness of a soap film in air.

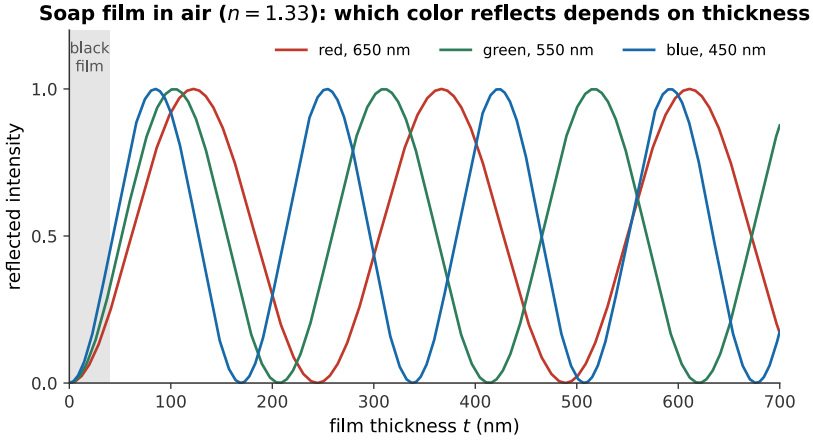

Figure 4.13: Reflected intensity from a soap film in air ($n = 1.33$) versus film thickness, for three visible wavelengths. Different colors peak at different thicknesses, which is why a draining soap film shows moving bands of color. As $t \to 0$ all three curves go to zero: the film turns black just before it bursts. Generated with matplotlib; see `scripts/figures/`.

Two features of the figure are worth dwelling on. First, the three curves peak at different thicknesses, so a soap film whose thickness varies from place to place — as it always does, since gravity drains it — shows bands of color that drift downward as the film thins. Second, and more striking: as $t \to 0$ *all* the curves go to zero. A film very much thinner than $\lambda/4n_{\text{film}}$ contributes almost no path difference, so the two reflections are left with only the π shift and cancel for every visible wavelength at once. The film goes **black** — a genuinely counterintuitive prediction, and one you can verify by watching a soap film held vertically in a wire loop: a black band appears at the top, spreads downward, and moments later the film bursts. Newton described this effect in the 1670s, without being able to explain it.

Worked Example: The Color of a Soap Film

A soap film ($n = 1.33$) in air is 100 nm thick. Which visible wavelength does it reflect most strongly, viewed at normal incidence?

This is Case A (air–film–air, so only the front reflection flips), and constructive reflection requires $2n_{\text{film}}t = (m + \frac{1}{2})\lambda$. Solving for λ:

$$\lambda = \frac{2n_{\text{film}}t}{m + \frac{1}{2}}. \tag{4.40}$$

For $m = 0$: $\lambda = 2(1.33)(100\text{ nm})/0.5 = 532$ nm — green. For $m = 1$: $\lambda = (266\text{ nm})/1.5 = 177$ nm, deep in the ultraviolet and invisible. So this film looks green, and only the $m = 0$ order matters. That is characteristic of thin films: they are thin enough that only the lowest one or two orders land in the visible, which is why their colors are broad and pastel rather than a sequence of sharp spectral lines.

Worked Example: An Antireflection Coating

A camera lens ($n_{\text{glass}} = 1.52$) is coated with magnesium fluoride ($n_{\text{MgF}_2} = 1.38$) to minimize reflection at $\lambda = 550$ nm, the middle of the visible spectrum where the eye is most sensitive. Light is incident from air. Find the minimum coating thickness.

Apply the recipe. Front reflection: air (1.00) into MgF_2 (1.38) — index increases, so π. Back reflection: MgF_2 (1.38) into glass (1.52) — index increases again, so π. Both flip: this is **Case B**, net shift 0, and destructive reflection requires the odd-quarter-wave condition

$$2n_{\text{MgF}_2}t = \left(m + \tfrac{1}{2}\right)\lambda. \tag{4.41}$$

The thinnest coating uses $m = 0$:

$$t = \frac{\lambda}{4n_{\text{MgF}_2}} = \frac{550\text{ nm}}{4(1.38)} = 99.6\text{ nm} \approx 100\text{ nm}, \tag{4.42}$$

the standard "quarter-wave" coating. Two remarks. First, the cancellation is not perfect even at 550 nm, because the two reflected beams have slightly different amplitudes; complete cancellation would require $n_{\text{coating}} = \sqrt{n_{\text{air}}n_{\text{glass}}} = 1.23$, and MgF_2 at 1.38 is simply the closest durable material available. Second, the condition is wavelength-specific: at 450 nm and 700 nm the same coating still suppresses reflection, but only partially. The residual reflection is therefore richer in blue and red than in green, which is exactly the faint purple sheen you see on a coated lens.

Wedges and Newton's Rings

A film need not have uniform thickness. Press two flat glass plates together and separate one edge with a thin spacer — a hair, a wire, a sheet of paper — and the air gap between them forms a **wedge** whose thickness grows linearly along the plates. This is Case A: at the glass-to-air surface the index *decreases*, so there is no shift, while at the air-to-glass surface below it increases, so that reflection flips. Exactly one flip, net shift π, and with $n_{\text{film}} = 1$ for air the dark fringes fall where $2t = m\lambda$ — including at $t = 0$, so the line of contact is dark. That dark contact line is a useful check that you have assigned the case correctly.

If the wedge has thickness D at a distance L from the contact line, then $t = Dx/L$ at position x, and successive dark fringes are separated by

$$\Delta x = \frac{\lambda L}{2D}. \tag{4.43}$$

Counting the fringes over the whole length gives $2D/\lambda$, so the wedge is a way of measuring a very small thickness by counting a large number.

Replace the top plate with a slightly convex lens resting on the flat and the fringes become concentric circles: **Newton's rings**. For a lens of radius of curvature R, geometry gives the gap at radius r as $t \approx r^2/2R$, so the dark rings fall at

$$r_m = \sqrt{m\lambda R}, \qquad m = 0, 1, 2, \ldots \tag{4.44}$$

with a dark spot at the center. The rings crowd together as m grows, since $r_m \propto \sqrt{m}$. Newton's rings remain a standard optical-shop test: any departure of the ring pattern from perfect circles reveals a departure of the surface from a perfect sphere, at a sensitivity of a fraction of a wavelength.

> **Warning**
> The bright/dark conditions derived in this section describe light reflected back toward the source. Viewed in transmission instead — looking through the wedge or the lens rather than at its reflection — the same film shows the *complementary* pattern: wherever the reflected light is dark, the transmitted light is bright, and vice versa. This is not a separate phase-shift rule to memorize; it follows from energy conservation in a film that absorbs essentially nothing, so whatever fraction of the light is not reflected must be transmitted. A dark spot at the center of Newton's rings in reflection is therefore a bright spot at the center in transmission.

Worked Example: Measuring a Wire with an Air Wedge

Two flat glass plates 10.0 cm long are in contact at one end and separated at the other by a thin wire. Illuminated from above with sodium light ($\lambda = 589$ nm), the plates show 170 dark fringes between the contact line and the wire. Find the wire's diameter.

The wire's diameter is the wedge thickness D at the far end. Dark fringes occur wherever $2t = m\lambda$, so the fringe count from $t = 0$ to $t = D$ is $m_{\max} = 2D/\lambda$:

$$D = \frac{m_{\max}\lambda}{2} = \frac{170\,(589 \times 10^{-9}\ \text{m})}{2} = 5.01 \times 10^{-5}\ \text{m} = 50.1\ \mu\text{m}. \tag{4.45}$$

The plate length never entered — it only sets the fringe *spacing*, $\Delta x = \lambda L/2D = 0.588$ mm, which is what makes the fringes countable by eye. A 50 μm wire has been measured to within a fraction of a micrometer using nothing but two pieces of glass, a sodium lamp, and patience.

The Michelson Interferometer

The **Michelson interferometer** (Figure 4.14) is Young's two-path experiment rebuilt with mirrors, and it is the most consequential single instrument in this book.

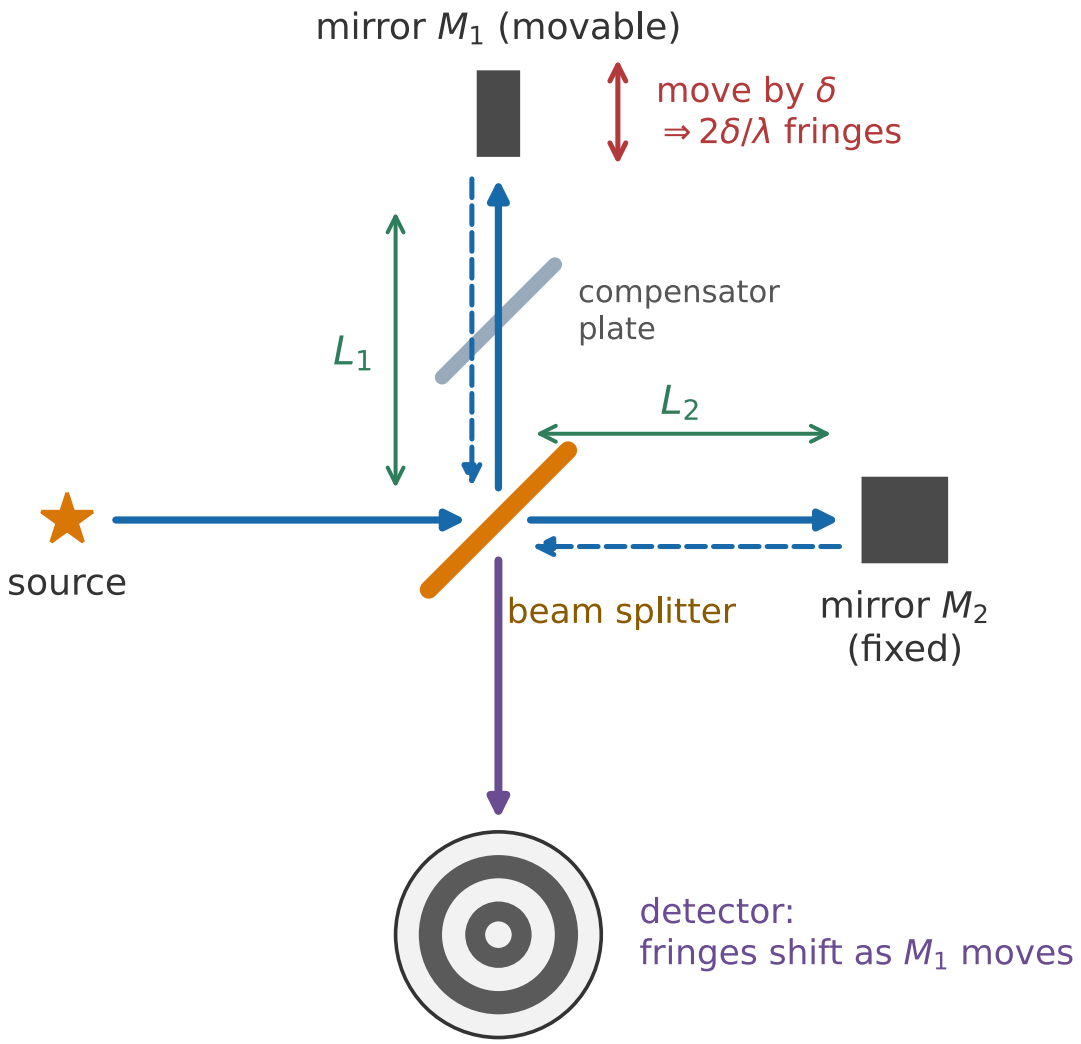

Figure 4.14: The Michelson interferometer. A beam splitter divides the incoming light into two perpendicular arms; each returns from a mirror and the two recombine at the splitter. Moving one mirror by δ changes that arm's round trip by 2δ and sweeps $2\delta/\lambda$ fringes past the detector. The compensator plate equalizes the amount of glass traversed by the two beams. Original schematic generated with matplotlib; see `scripts/figures/`.

A beam splitter — a half-silvered mirror — divides an incoming beam into two perpendic-

ular paths of lengths L_1 and L_2. Each path ends at a mirror that sends the light back, and the beam splitter recombines the two returning beams and sends them to a detector. The round-trip path difference is $2(L_1 - L_2)$, so the recombined beams interfere according to

$$\Delta r = 2(L_1 - L_2). \tag{4.46}$$

> A "half-silvered" beam splitter need not split the light exactly 50/50 in a real instrument. Nothing in the fringe-counting analysis of this section depends on the split ratio — only on the fact that both paths carry some of the original beam and recombine coherently at the end.

Now translate one mirror through a distance δ. That arm's round trip changes by 2δ, and the fringe pattern shifts by

$$\Delta N = \frac{2\delta}{\lambda} \qquad \text{fringes.} \tag{4.47}$$

One full fringe passes the detector for every *half* wavelength of mirror motion. Since $\lambda/2 \approx 300$ nm for visible light, and fringe positions can be interpolated to a small fraction of a fringe, the instrument measures displacements of a few nanometers. That is the whole reason for its importance: it converts a length comparison into a fringe count, and light's wavelength is a very fine ruler.

The small tilted plate in Figure 4.14 is the **compensator**. Without it, the beam going to M_2 passes through the glass of the beam splitter three times while the beam going to M_1 passes through it once — a large and, worse, wavelength-dependent optical path difference. The compensator is an identical piece of uncoated glass placed in the other arm so that both beams traverse the same thickness. It is unnecessary with a laser but essential with white light.

Both arms, and the fringes they produce, can be manipulated directly in Figure 4.15. Translating one mirror sweeps the fringe count given by Equation (4.47); shortening the coherence length washes the fringes out, which is the coherence-length constraint of the earlier section made visible.

Worked Example: Counting Fringes

In a Michelson interferometer illuminated with sodium light, $\lambda = 589$ nm, one mirror is slowly translated and 1200 fringes are counted passing a reference mark. How far did the mirror move?

$$\delta = \frac{\Delta N\,\lambda}{2} = \frac{1200\,(589 \times 10^{-9}\ \text{m})}{2} = 3.53 \times 10^{-4}\ \text{m} = 0.353\ \text{mm}. \tag{4.48}$$

Note the leverage: a third of a millimeter of motion, resolved into 1200 countable events.

Worked Example: The Refractive Index of Air

A transparent cell of length $L = 5.00$ cm is placed in one arm and slowly evacuated. As the air is pumped out, 49.7 fringes are counted. Find the refractive index of air at $\lambda = 589$ nm.

The light crosses the cell twice, so removing air of index n changes the optical path by $2L(n-1)$, and the fringe count is

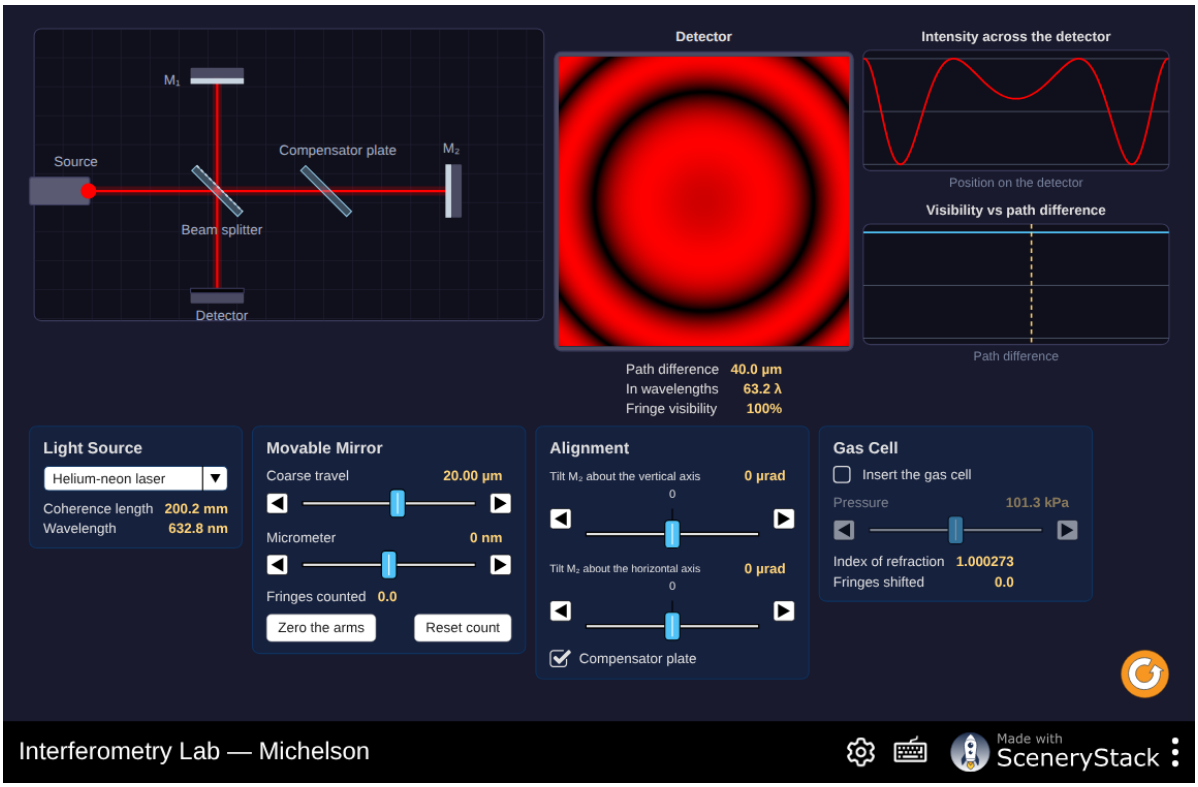

Figure 4.15: A physical-optics model of the Michelson interferometer, together with the Mach–Zehnder and Fabry–Pérot geometries. Move a mirror and count fringes; change the source's coherence length and watch the visibility collapse. Interactive simulation: Interferometry Lab

$$\Delta N = \frac{2L(n-1)}{\lambda} \quad \Longrightarrow \quad n - 1 = \frac{\Delta N\,\lambda}{2L} = \frac{49.7\,(589 \times 10^{-9}\ \text{m})}{2(0.0500\ \text{m})} = 2.93 \times 10^{-4}, \tag{4.49}$$

so $n = 1.000293$ — the accepted value for dry air at standard conditions. The interferometer has measured a refractive index that differs from unity in the fourth decimal place, from a fringe count that a student can do by eye.

What Interferometers Are For

- **The Michelson–Morley experiment.** As described in Chapter 1, rotating the apparatus should have exchanged the roles of the arm parallel to Earth's motion through the ether and the arm perpendicular to it, shifting the fringes by about 0.4 fringe against a sensitivity of 0.01 fringe. No shift was ever seen, at any orientation or any time of year. The fringe-counting relation in Equation (4.47) is exactly what converted a null optical measurement into a decisive statement about the structure of spacetime.
- **Fourier-transform spectroscopy.** Record the detector signal as a function of mirror position and you have the *autocorrelation* of the light; its Fourier transform is the spectrum. Nearly every infrared spectrometer in a modern chemistry laboratory is a Michelson interferometer operated this way.
- **Metrology.** Surface flatness, machine-tool calibration, and — until the 1983 redefinition of the meter in terms of c — the international length standard itself.
- **Gravitational-wave detection.** LIGO and Virgo are Michelson interferometers with 4 km arms, folded optically to an effective length of hundreds of kilometers, measuring mirror displacements of order 10^{-19} m — about one ten-thousandth of a proton's diameter. The first detection, in September 2015, came from two merging black holes 1.3 billion light-years away. The instrument that failed to find the ether became the instrument that found the ripples in spacetime that replaced it.

Looking Ahead: Interference One Photon at a Time

Everything in this chapter treats light as a classical wave, and the treatment works. It is worth flagging now, though, that the story does not end here.

> **See Also**
> Particle Properties of Waves picks up the single-photon result previewed here, developing the photon concept and the photoelectric effect that first forced physicists to take a particle picture of light seriously again.

Turn the source in Young's experiment down. Not a little — down until the light is so faint that, on average, only one quantum of light is inside the apparatus at any moment. A 1 mW helium–neon laser emits about 3×10^{15} photons per second; light crosses a meter-long apparatus in about 3 ns. A rate below roughly 10^{8} photons per second keeps the mean number of photons in the apparatus below about 0.3; individual emission intervals are still random, but simultaneous photons are then uncommon. G. I. Taylor performed the experiment in 1909 using an attenuated gas flame and a three-month exposure, and it has been repeated countless times since with single-photon sources and imaging detectors.

The result is that the screen records individual, localized hits — dots, one at a time, apparently at random. But as the dots accumulate over hours, they build up precisely the $\cos^2$ fringe pattern derived in this chapter. The probability amplitude associated with each photon interferes between the two paths; there is no second photon for it to interfere with. That fact cannot be accommodated by any picture in which the photon simply goes through one slit or the other, and it is the central puzzle that Chapters 6 and 7 take up.

For now, the wave description stands on its own, and the next chapter completes it.

4.5 Summary

- **Light is an electromagnetic wave.** Detectors measure the time-averaged intensity, $I \propto E_0^2$, so *amplitudes* superpose and the intensity is the square of the sum — never the sum of the intensities. In a medium of index n, the wavelength shrinks to λ/n, and what governs interference is the **optical path length** nd.
- **Wave effects hide behind a small number.** The only dimensionless quantity available is λ/a, and for everyday apertures it is $\sim 10^{-4}$. Ray optics is wave optics in the limit $\lambda/a \to 0$ — an instance of the correspondence principle.
- **Huygens's principle** treats every point of a wavefront as a source of secondary wavelets whose envelope is the later wavefront; it predicts that waves spread after an aperture, by an amount governed by λ/a.
- **Path difference becomes phase difference** through $\phi = (2\pi/\lambda)\Delta r$. Constructive interference requires $\Delta r = m\lambda$, destructive requires $\Delta r = (m + \frac{1}{2})\lambda$.
- **Young's double slit** gives bright fringes at $d\sin\theta = m\lambda$ and dark fringes at $d\sin\theta = (m + \frac{1}{2})\lambda$; for $L \gg d$ and small angles, the fringes are evenly spaced with $\Delta y = \lambda L/d$. Only orders with $|m| \le d/\lambda$ exist.
- **Coherence** is required for a stable pattern. Real sources emit short wave trains with randomly resetting phase; the **coherence time** τ_c and **coherence length** $\ell_c = c\tau_c \approx \lambda^2/\Delta\lambda$ set the largest usable path difference, and $\lambda/\Delta\lambda$ estimates the number of visible fringes. Spatial coherence — supplied by Young's first slit, or built into a laser — is what lets two slits act as one source.
- **Phasor addition** gives the two-slit intensity $I = I_0\cos^2(\pi d\sin\theta/\lambda)$ with $I_0 = 4I_1$. The factor of four is offset by the zeros: the pattern *redistributes* energy, averaging back to $2I_1$.
- N **equally spaced slits** give $I = I_1[\sin(N\phi/2)/\sin(\phi/2)]^2$. Principal maxima stay at $d\sin\theta = m\lambda$, rise as N^2, and narrow as $1/N$, with $N-2$ weak secondary maxima between them — the basis of the diffraction grating (Chapter 5).
- **Thin films** interfere via the extra optical path $2n_{\text{film}}t$ plus any π shift from reflection at a boundary where the index increases. Count the shifts first: one flip gives constructive reflection at $2n_{\text{film}}t = (m + \frac{1}{2})\lambda$; zero or two flips reverse the conditions. This explains soap-film colors, the black film, quarter-wave antireflection coatings, wedge fringes, and Newton's rings ($r_m = \sqrt{m\lambda R}$).
- **The Michelson interferometer** converts a mirror displacement δ into a fringe count $\Delta N = 2\delta/\lambda$, giving nanometer sensitivity. It produced the null result of Chapter 1, and, scaled to kilometers, the first detection of gravitational waves.

4.6 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens on a double-slit intensity pattern.

1. **Multiple choice, from a figure.** If the slit separation d increases while the wavelength and screen distance stay fixed, what happens to the double-slit fringe spacing? Figure description: Double-slit intensity pattern with regularly spaced bright fringes. Choices: (A) It decreases because the spacing is proportional to 1/d. (B) It increases because the spacing is proportional to d. (C) It stays unchanged because only wavelength controls the pattern. (D) The bright and dark fringes exchange positions without changing spacing.

2. **True or false.** Reflection from a boundary leading into a higher-index medium adds a phase change of π.

3. **Drag the words.** Complete the interference conditions. A path difference Δr produces phase difference _____. Constructive interference occurs at _____, while destructive interference occurs at _____. Stable fringes require _____. (Terms: $2\pi\Delta r/\lambda$; $\Delta r = m\lambda$; $\Delta r =$ (m + 1/2)λ; coherence; $\Delta r = \lambda/m$; incoherence.)

4. **Fill in the blanks.** Complete the path-to-phase and fringe-spacing results. A path difference of $\lambda/2$ gives a phase difference of _____ radians, and doubling d makes the fringe spacing _____ as large.

5. **Mark the words.** Mark the two statements that explain coherence and energy conservation in interference. "A stable pattern requires a phase difference that remains predictable during observation. Interference redistributes energy between bright and dark regions; it does not create energy."

4.7 Conceptual Questions

1. Two flashlights are aimed at the same spot on a wall. Explain, in terms of coherence time, why no interference fringes appear, even though two light waves are certainly overlapping there.

2. In Young's experiment, what happens to the fringe spacing if (a) the slit separation d is doubled, (b) the screen distance L is doubled, (c) the light is changed from red to blue, (d) one slit is covered? Answer each in one sentence.

3. The central maximum in a double-slit pattern has four times the intensity that one slit alone would produce. Explain why this does not violate conservation of energy, and state where the extra energy comes from.

4. Explain why the $m = 0$ fringe is white when a double slit is illuminated with white light, while every other fringe is colored.

5. A soap film held vertically in a wire loop develops a black band at the top just before it bursts. Explain why the film appears black there rather than showing some color, and why the band appears at the *top*.

6. A camera lens with an antireflection coating still shows a faint purple reflection. Explain why the coating cannot eliminate reflection at all visible wavelengths at once.

7. Two glass plates in contact at one edge form an air wedge. Is the fringe at the line of contact bright or dark? Justify your answer by counting phase shifts on reflection, and explain what you would conclude if you observed the opposite.

8. Explain why a Michelson interferometer with arms differing by several centimeters shows fringes with a laser but not with a sodium lamp, even though the sodium lamp is far from white.

4.8 Problems

Exercise 4.1
(Straightforward)

Light of wavelength 633 nm (a helium–neon laser) illuminates two slits separated by $d = 0.120$ mm. Find the angle θ to (a) the third bright fringe and (b) the second dark fringe, measured from the central axis.

Exercise 4.2
(Straightforward)

In a double-slit experiment with $d = 0.250$ mm and a screen at $L = 1.40$ m, bright fringes are observed with spacing $\Delta y = 3.30$ mm. Find the wavelength of the light used, and identify approximately its color.

Exercise 4.3
(Straightforward)

A double slit with $d = 0.0800$ mm is illuminated at $\lambda = 480$ nm with the screen 2.50 m away. (a) Find the fringe spacing. (b) How many bright fringes fall on a screen 8.0 cm wide, centered on the axis? (c) What is the largest order that exists in principle?

Exercise 4.4
(Straightforward)

Two identical in-step sources emit light of wavelength 620 nm. Find the smallest nonzero path difference that produces (a) a bright point, (b) a dark point, (c) a point whose intensity is half the maximum.

Exercise 4.5
(Straightforward)

The double slit of Problem 3 is immersed, along with its source and screen, in a liquid of refractive index $n = 1.47$. Find the new fringe spacing, and explain in one sentence why the answer does not depend on where in the apparatus the liquid is.

Exercise 4.6
(Straightforward)

A source has a central wavelength of 600 nm and a wavelength spread $\Delta\lambda = 0.02$ nm, typical of a good single-mode laser. (a) Estimate its coherence length. (b) Estimate the number of fringes visible before the pattern washes out. (c) Repeat both estimates for white light, $\lambda \approx 550$ nm with $\Delta\lambda \approx 300$ nm, and comment on why Young's original experiment showed only a handful of fringes.

Exercise 4.7
(Straightforward)

Starting from $I = I_0 \cos^2(\phi/2)$ with $\phi = (2\pi/\lambda) d \sin\theta$, show that, in the small-angle approximation, the full width at half maximum of each fringe is exactly half the fringe spacing, independent of λ, d, and L.

Exercise 4.8
(Straightforward)

Using the N-slit intensity formula, evaluate $I(\theta = 0)$ for $N = 2$ and $N = 4$ slits with the same single-slit intensity I_1, and verify the scaling $I_{\max} \propto N^2$. Then show that the total light delivered by N slits is proportional to N, not N^2, by combining the peak height with the peak width.

Exercise 4.9
(Moderate)

Six equally spaced slits with $d = 3.00$ μm are illuminated at $\lambda = 600$ nm. (a) Find $\sin\theta$ for all the principal maxima. (b) Find the directions of the zeros between the central and first-order maxima. (c) How many secondary maxima lie between them?

Exercise 4.10
(Moderate)

A soap film ($n = 1.33$) in air is viewed at near-normal incidence with $\lambda = 500$ nm light. (a) Explain, by counting reflections, why only one of the two picks up a π shift. (b) Find the two smallest nonzero thicknesses at which the film appears bright in reflection. (c) Find the smallest nonzero thickness at which it appears dark.

Exercise 4.11
(Moderate)

A thin layer of oil ($n = 1.45$) floats on water ($n = 1.33$). (a) Determine whether this is Case A or Case B in the classification of this chapter. (b) Find the minimum oil thickness that reflects $\lambda = 600$ nm light strongly. (c) Repeat parts (a) and (b) for the same oil film resting instead on a substrate of index 1.60, and explain why the answer changes by a factor of two.

Exercise 4.12
(Moderate)

Design an antireflection coating of magnesium fluoride ($n = 1.38$) on glass ($n = 1.52$) that minimizes reflection at $\lambda = 600$ nm in air. (a) Find the minimum thickness. (b) Compute $2n_{\text{coating}}t/\lambda$ at 450 nm and at 700 nm, and use the results to explain the residual purple reflection of coated optics.

Exercise 4.13
(Moderate)

A plano-convex lens with radius of curvature $R = 2.00$ m rests on a flat glass plate and is illuminated from above with $\lambda = 589$ nm light. (a) Find the radius of the tenth dark ring. (b) Explain why the center of the pattern is dark. (c) Explain why the rings crowd closer together as m increases.

Exercise 4.14
(Moderate)

Two flat plates 15.0 cm long touch at one end and are separated at the other by a sheet of foil. Illuminated at $\lambda = 546$ nm, the plates show dark fringes spaced 0.750 mm apart. Find the thickness of the foil.

Exercise 4.15
(Moderate)

In a Michelson interferometer illuminated at $\lambda = 546$ nm, one mirror is translated by 0.200 mm. How many fringes pass a reference mark?

Exercise 4.16
(Moderate)

A gas cell 8.00 cm long in one arm of a Michelson interferometer is filled with carbon dioxide at atmospheric pressure, and 122 fringes are counted as it fills. Find the refractive index of carbon dioxide at $\lambda = 589$ nm.

Exercise 4.17
(Challenging)

Explain, using the concept of optical path length, why the thin-film conditions use $2n_{\text{film}}t$ rather than $2t$, and estimate the error you would make in the predicted color of a 100 nm soap film if you dropped the factor of n.

Exercise 4.18
(Challenging)

Revisit the Michelson–Morley experiment of Chapter 1 using this chapter's fringe-shift relation. The predicted ether-wind path difference between the two arms was of order Lv^2/c^2 for arm length L and Earth's orbital speed $v \approx 3 \times 10^4$ m/s. Estimate the expected fringe shift for $L = 11$ m and $\lambda = 590$ nm, and compare it with the experiment's sensitivity of about 0.01 fringe.

Exercise 4.19
(Challenging)

LIGO measures mirror displacements of about 10^{-19} m using $\lambda = 1064$ nm light. (a) Using $\Delta N = 2\delta/\lambda$, find the corresponding fringe shift. (b) Your answer should be an absurdly small fraction of a fringe; explain what this implies about the additional techniques (optical cavities that fold the beam path, and averaging over enormous photon numbers) that a real detector must use.

Exercise 4.20

(Challenging)

A double slit is illuminated with light containing two wavelengths, 480 nm and 600 nm, with $d = 0.150$ mm and $L = 2.00$ m. Find the smallest nonzero distance from the center of the screen at which a bright fringe of one wavelength coincides exactly with a bright fringe of the other.

Chapter 5

Diffraction of Light

Learning Objectives

By the end of this chapter, you should be able to:

- Explain diffraction as the Huygens–Fresnel principle applied to a continuum of sources across a finite aperture, and say precisely how it differs — and does not differ — from the interference of Chapter 4.
- Distinguish Fresnel (near-field) from Fraunhofer (far-field) diffraction, and decide which applies in a given geometry.
- Derive the single-slit minimum condition by the pairwise-cancellation argument, and explain why $m = 0$ is excluded.
- Sketch and use the single-slit intensity pattern, including the width of the central maximum and the rapid decline of the secondary maxima.
- Explain how a real double slit's pattern is the product of a two-slit interference factor and a single-slit envelope, and locate missing orders.
- Apply the grating equation, compute angular dispersion, and use the resolving power $R = mN$ in spectroscopic problems.
- Apply the Rayleigh criterion for a circular aperture to telescopes, microscopes, and the eye, and distinguish the diffraction limit from practical limits such as atmospheric seeing.
- State Bragg's law, use it to extract interplanar spacings from measured X-ray diffraction angles, and explain why the same argument applies to electron and neutron beams.
- Describe how a hologram records and reconstructs a wavefront, and explain why it requires a coherent source.
- Use single-slit diffraction to make a first, order-of-magnitude argument for the uncertainty principle of Chapter 7.

Introduction

In 1818 the French Academy of Sciences held a prize competition on the nature of light. Augustin-Jean Fresnel submitted a wave theory. Siméon Denis Poisson, a member of the judging committee and a convinced supporter of Newton's particle theory, worked through Fresnel's mathematics and triumphantly produced what he took to be its refutation: if Fresnel were right, then the shadow of a small circular disk, illuminated by a point source, would have a *bright spot at its exact center* — light would diffract around the rim, arrive at the axis having traveled equal distances from every point of that rim, and interfere constructively. The prediction was patently ridiculous.

A member of that committee, François Arago, went and looked. The spot was there.

That bright point — now called the Poisson spot, or the Arago spot, depending on whose role in the story one prefers to emphasize — is a fair emblem for this chapter. **Diffraction** is what waves do when they encounter an obstacle or a finite aperture: they bend into the geometric shadow, and they interfere with themselves in the process, producing patterns of light and dark that no ray-tracing argument can generate.

Chapter 4 worked with idealized slits: point-like Huygens sources, infinitesimally narrow, contributing one wavelet apiece. A real slit has a finite width a, and *every* point across that width emits a wavelet. Summing those contributions is the whole subject of this chapter. Three results follow, in increasing order of consequence. First, a single slit, all by itself, produces a pattern of bright and dark bands. Second, real multiple-slit devices — above all the diffraction grating — combine the sharp interference maxima of Chapter 4 with a diffraction envelope that decides how bright each of those maxima actually is. Third, and most importantly, because every real optical instrument gathers light through an aperture of finite size, diffraction sets an absolute limit on the fineness of detail any instrument can resolve — a limit that no improvement in polishing, alignment, or manufacture can evade, because it is imposed by the wavelength of light itself.

5.1 Diffraction Basics and the Single Slit

Diffraction Versus Interference: A Note on Words

It is worth settling a piece of vocabulary that causes more confusion than it should.

There is no physical distinction between interference and diffraction. Both are the same thing: coherent waves from different places arriving at a common point and adding as fields. The words differ only in bookkeeping convention. When the sources are a small, discrete set — two slits, five slits — the custom is to say *interference* and to write a sum. When the sources form a continuum across an aperture, the custom is to say *diffraction* and to write an integral. Richard Feynman put it bluntly: "no one has ever been able to define the difference between interference and diffraction satisfactorily. It is just a question of usage, and there is no specific, important physical difference between them."

The reason the distinction survives is practical. In a real double slit there are two length scales — the slit separation d and the slit width a, with $d > a$ — and they produce structure on two very different angular scales. Calling the fine structure "interference" and the coarse envelope "diffraction" is a convenient way to keep track of which is which. That is all it is.

The Huygens–Fresnel Principle

Fresnel's contribution was to make Huygens's construction quantitative. Where Huygens drew an envelope of wavelets, Fresnel insisted on *adding the wavelets as waves, with their phases*, and then squaring the result to get an intensity. Stated for an aperture:

> Every point of an open aperture acts as a source of secondary wavelets of equal amplitude and of the same phase as the incoming wave at that point. The field at any observation point is the sum, over the whole aperture, of these wavelets, each with the phase it has accumulated on its journey.

This **Huygens–Fresnel principle** reproduces the results of Chapter 4 when the aperture is a set of very narrow slits, and everything in this chapter when it is not. The sum becomes an integral, but the physics — add fields, account for phase, then square — is unchanged.

Near Field and Far Field

Whether the sum is easy or hard depends on the geometry, and the distinction matters enough to have names.

If the screen is far enough away that the rays converging on any observation point are effectively parallel, the phase of each wavelet is a *linear* function of its position in the aperture, and the integral is elementary. This is **Fraunhofer diffraction**, or far-field diffraction, and it is the case treated throughout this chapter.

If the screen is close, the rays are noticeably non-parallel, the phase varies *quadratically* across the aperture, and the pattern depends in a complicated way on distance. This is **Fresnel diffraction**, or near-field diffraction. The Poisson spot of the introduction is a Fresnel phenomenon; so are the fine bright and dark bands you can see just inside the edge of a sharp shadow cast by a small bright source.

The dividing line is set by the **Fresnel number**

$$F = \frac{a^2}{L\lambda}, \tag{5.1}$$

> F is dimensionless — a^2 and $L\lambda$ are both areas — so it can be compared directly to 1 regardless of what length unit a, L, and λ happen to be quoted in, as long as all three use the same one.

where a is the aperture size and L the aperture-to-screen distance. Fraunhofer diffraction applies when $F \ll 1$; Fresnel diffraction when $F \gtrsim 1$. For a 0.05 mm slit, 550 nm light, and a screen 2 m away, $F = (5\times10^{-5}\ \text{m})^2/[(2\ \text{m})(5.5\times10^{-7}\ \text{m})] = 2\times10^{-3}$, comfortably in the far field. Conveniently, a lens converts the far field into something you can put on a bench: any lens placed after the aperture brings the parallel bundles to a focus in its focal plane, so the Fraunhofer pattern appears there no matter how short the bench.

> **Note**
> Joseph von Fraunhofer, the optician whose name now labels the entire far-field regime, made his most consequential discovery almost as a side effect. Testing prism spectrometers around 1814, he noticed that the Sun's spectrum was crossed by hundreds of sharp dark lines at fixed wavelengths — the Fraunhofer lines — and went on to rule some of the first precision diffraction gratings by hand, specifically to measure those wavelengths accurately. Decades later, Gustav Kirchhoff and Robert Bunsen explained the lines as absorption by specific elements in the Sun's outer atmosphere, founding the science of spectroscopy that the diffraction grating, later in this chapter, still serves.

Single-Slit Diffraction

Consider a slit of width a illuminated by a plane wave of wavelength λ, with a distant screen. The clever step, due to Fresnel, is not to attempt the whole sum at once but to ask a narrower question: *in which directions does everything cancel?*

Divide the slit conceptually into two equal halves, each of width $a/2$, as in Figure 5.1. Pair up the topmost point of the upper half with the topmost point of the lower half; they are a distance $a/2$ apart. Pair the second point of the upper half with the second point of the lower half; also $a/2$ apart. Every point in the upper half has a partner in the lower half exactly $a/2$ below it. If the light heading off at angle θ from each pair is exactly out of phase — path difference $\lambda/2$ — then *every* pair cancels, and so does the whole slit:

$$\frac{a}{2}\sin\theta = \frac{\lambda}{2} \qquad \Longrightarrow \qquad a\sin\theta = \lambda. \tag{5.2}$$

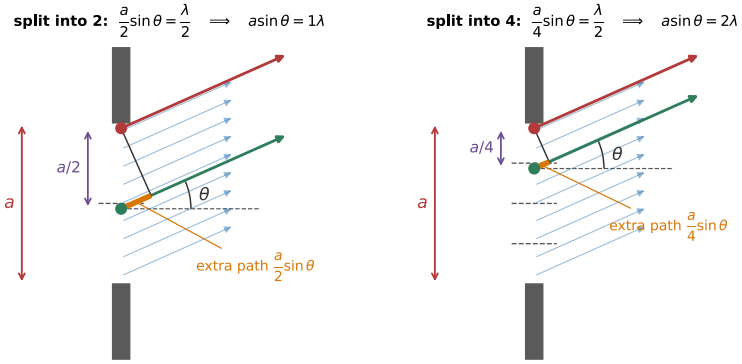

Figure 5.1: Canceling the slit against itself. Dividing the slit into two strips pairs each point with a partner $a/2$ away; when those pairs are a half wavelength out of step, the whole slit cancels, giving $a\sin\theta = \lambda$. Dividing into four strips pairs points $a/4$ apart and gives $a\sin\theta = 2\lambda$, and so on. Original schematic generated with matplotlib; see `scripts/figures/`.

Repeat the argument with four strips (Figure 5.1): now partners are $a/4$ apart, each pair cancels when $(a/4)\sin\theta = \lambda/2$, and the whole slit cancels at $a\sin\theta = 2\lambda$. With $2m$ strips the condition is $a\sin\theta = m\lambda$. The complete set of **single-slit minima** is therefore

$$a\sin\theta = m\lambda, \qquad m = \pm1, \pm2, \pm3, \ldots \qquad \text{(diffraction minima)}. \tag{5.3}$$

Two things about this formula regularly trip people up, and both are worth stating explicitly.

It locates minima, not maxima. Compare with the double-slit condition $d\sin\theta = m\lambda$, which locates *maxima*. The two formulas look almost identical and mean opposite things. The distinction is not a convention: it comes from the fact that the pairing argument can only ever demonstrate cancellation.

$m = 0$ **is excluded.** At $\theta = 0$ every point of the slit is in phase with every other, so the wavelets add rather than cancel — there is nothing to pair against. The center of the pattern is a broad, bright **central maximum**, not a minimum. Formally, the pairing argument demands a path difference of $\lambda/2$ within each pair, and $\theta = 0$ supplies a path difference of zero.

The Width of the Central Maximum

The central maximum is bounded by the first minima on either side, at $\sin\theta = \pm\lambda/a$. Its angular half-width is therefore

$$\theta_1 = \arcsin\frac{\lambda}{a} \approx \frac{\lambda}{a} \quad (\text{for } \lambda \ll a), \tag{5.4}$$

exactly the λ/a scaling that the dimensional argument of Chapter 4 demanded. Note that in $\sin\theta$ the central maximum runs from $-\lambda/a$ to $+\lambda/a$, while every other maximum spans only λ/a: the central peak is **twice as wide** as the rest, in addition to being far brighter.

The most important qualitative consequence is that *narrowing the slit widens the pattern*, as Figure 5.2 shows.

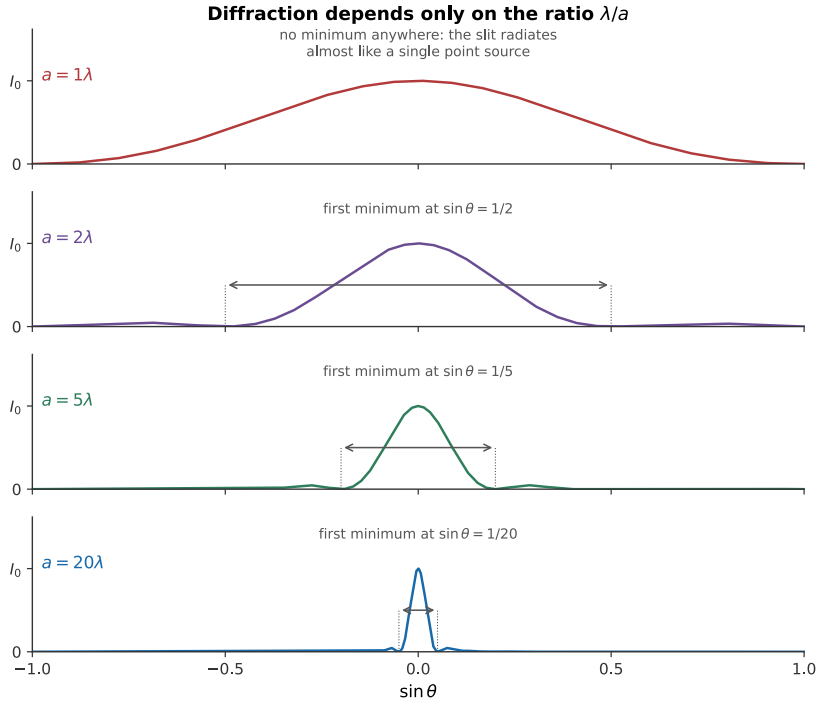

Figure 5.2: Single-slit patterns for four slit widths. The central maximum spans $\sin\theta = \pm\lambda/a$, so it narrows as the slit widens: a slit 20λ across concentrates the light into a beam, while a slit only 1λ across has no minimum at all and radiates almost like a single point source. Generated with matplotlib; see `scripts/figures/`.

This is the opposite of the naive expectation that a narrower opening should produce a narrower beam, and it is a genuinely wave-like signature. It also gives the correct picture of the limiting case: when $a \leq \lambda$, the condition $\sin\theta = \lambda/a \geq 1$ has no solution, there are no minima anywhere, and the slit radiates into the entire forward hemisphere — behaving, exactly as Chapter 4 assumed, like a single Huygens point source. That is the justification, after the fact, for treating the slits of the double-slit experiment as points.

The inverse relationship between the aperture and its pattern is worth seeing rather than reading about, and Figure 5.3 shows the two side by side: the aperture on the left, the far-field pattern it produces on the right, with the aperture's size and the wavelength both on sliders. Shrinking the opening spreads the pattern, and by exactly the ratio λ/a derived above. Two further controls make the point sharper than a single slit can. Squashing the round aperture into an ellipse produces an elliptical pattern elongated along the *other* axis, so the reciprocity holds direction by direction and not just on average; and the non-circular shapes on offer produce patterns of matching symmetry — an anticipation of the circular-aperture section below, where a round hole is found to give rings rather than a row of bands, and of the crystal lattices at the end of the chapter, where a periodic array of holes gives a periodic array of spots.

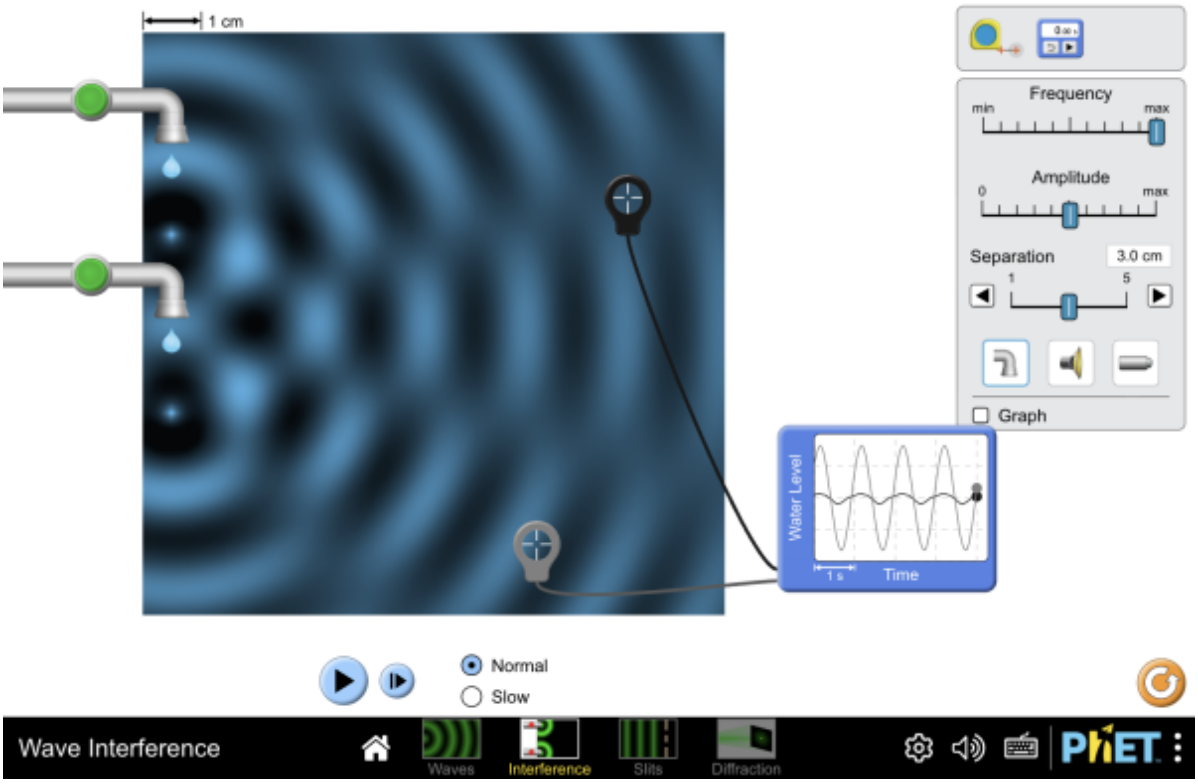

Figure 5.3: An aperture and the far-field diffraction pattern it produces, with aperture size, shape, and wavelength adjustable. Everything narrow in the aperture is wide in the pattern, and vice versa — the $\theta_1 \approx \lambda/a$ scaling of this section, seen whole rather than one angle at a time. Interactive simulation: Wave Interference

Worked Example: Width of the Central Maximum

A slit of width $a = 0.0400$ mm is illuminated with $\lambda = 580$ nm light, and the pattern is observed on a screen $L = 2.00$ m away. Find the width of the central maximum.

The first minima are at

$$\sin\theta_1 = \frac{\lambda}{a} = \frac{580 \times 10^{-9}\ \text{m}}{4.00 \times 10^{-5}\ \text{m}} = 0.0145, \tag{5.5}$$

so $\theta_1 = 0.0145$ rad $= 0.831°$ (the small-angle approximation is excellent here). On the screen,

$$y_1 = L\tan\theta_1 \approx L\theta_1 = (2.00\ \text{m})(0.0145\ \text{rad}) = 0.0290\ \text{m}, \tag{5.6}$$

and the central maximum, running from $-y_1$ to $+y_1$, has full width

$$w = 2y_1 = 5.80\ \text{cm}. \tag{5.7}$$

Check the trend: halving the slit width to 0.0200 mm would double this to 11.6 cm.

Worked Example: A Slit Too Narrow to Have Minima

At what slit width does the first minimum disappear entirely, for $\lambda = 580$ nm?

The first minimum requires $\sin\theta_1 = \lambda/a \leq 1$, so it exists only if $a \geq \lambda = 580$ nm. A slit narrower than one wavelength produces a smooth, minimum-free spread of light across the whole forward direction. Slits this narrow are difficult to make and transmit very little light, which is why the textbook idealization of a "point source slit" is easier to draw than to build — and why real double-slit experiments always show the envelope discussed below.

5.2 Intensity Patterns and the Double Slit

Intensity in Single-Slit Diffraction

The pairing argument gives the zeros; getting the full curve requires actually doing the sum. The phasor picture of Chapter 4 makes this almost graphical.

> **See Also**
> The phasor construction used below is the continuous limit of the discrete N-slit phasor sum from Chapter 4: there, finitely many equal-length phasors are added tip-to-tail one at a time; here, infinitely many infinitesimal ones curl smoothly into an arc. Working through that chapter's phasor derivation first makes the "chord of a circular arc" argument below feel like the natural next step rather than a new trick.

Divide the slit into N narrow strips and let $N \to \infty$. Each strip contributes a tiny phasor of the same length, and each is rotated slightly relative to its neighbor, because it sits slightly farther along the slit and so contributes a slightly different path. Let β be the *total* phase difference between the wavelet from one edge of the slit and the wavelet from the other:

$$\beta \equiv \frac{2\pi}{\lambda}\, a \sin\theta. \tag{5.8}$$

The chain of tiny phasors therefore turns through a total angle β from beginning to end. A chain of equal segments turning at a uniform rate is an arc of a circle, and the resultant field is its **chord** (Figure 5.4).

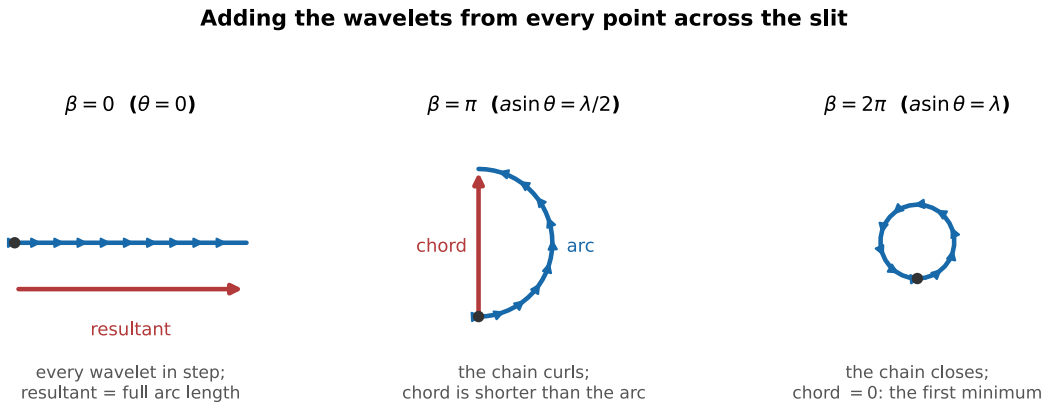

Figure 5.4: The single-slit phasor construction. The wavelets from across the slit form a chain of equal phasors turning through a total angle β. At $\theta = 0$ the chain is straight and the resultant equals the full arc length. As θ grows the chain curls and the chord falls behind the arc. At $\beta = 2\pi$ the chain closes into a circle and the resultant vanishes — the first minimum, $a \sin\theta = \lambda$. Original schematic generated with matplotlib; see `scripts/figures/`.

Everything follows from the geometry of a circular arc. Let the arc length be E_0 (this is the resultant at $\theta = 0$, when the chain is straight). An arc of length E_0 subtending angle β has radius $R = E_0/\beta$, and the chord across it is $2R\sin(\beta/2)$. So

$$E = 2\frac{E_0}{\beta}\sin\frac{\beta}{2} = E_0\,\frac{\sin(\beta/2)}{\beta/2}, \tag{5.9}$$

and squaring gives the **single-slit intensity pattern**

$$I(\theta) = I_0\left[\frac{\sin(\beta/2)}{\beta/2}\right]^2, \qquad \beta = \frac{2\pi}{\lambda}a\sin\theta. \tag{5.10}$$

The quantity in brackets is $\mathrm{sinc}(\beta/2)$, where $\mathrm{sinc}(x) \equiv \sin x/x$, so the pattern is often written $I = I_0\,\mathrm{sinc}^2(\beta/2)$.

> Two conventions for sinc circulate. The one used here, $\mathrm{sinc}(x) \equiv \sin x/x$, has its first zero at $x = \pi$. An engineering convention, common in signal-processing texts, instead defines $\mathrm{sinc}(x) \equiv \sin(\pi x)/(\pi x)$, with its first zero at $x = 1$. Check which one a source is using before comparing formulas across it.

Three checks confirm this is the right answer:

- At $\theta = 0$, $\beta \to 0$ and $\sin(\beta/2)/(\beta/2) \to 1$, giving $I = I_0$: the central maximum.
- $I = 0$ whenever $\beta/2 = m\pi$ with m a nonzero integer, i.e. $a\sin\theta = m\lambda$ — precisely the pairing-argument minima. The case $m = 0$ is excluded automatically, because $\sin x/x \to 1$ rather than 0 there.
- The chord can never exceed the arc, so $I \le I_0$ everywhere.

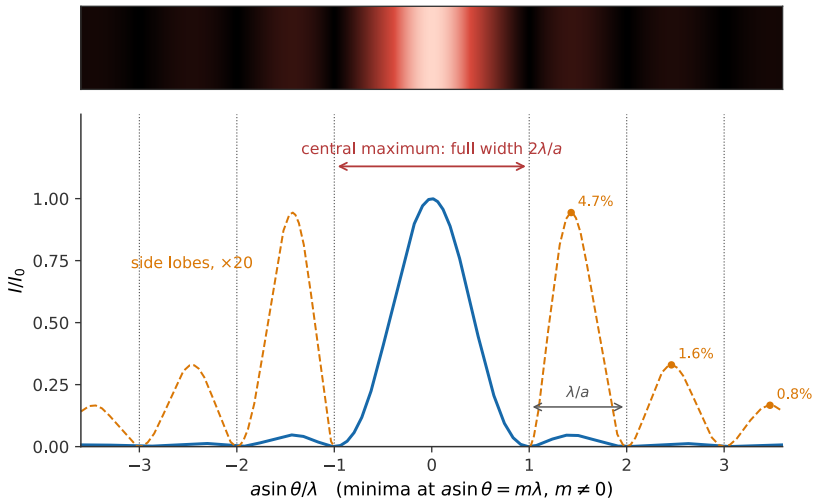

Figure 5.5: Single-slit diffraction intensity. The central maximum is twice as wide as the side lobes and vastly brighter — the first side lobe reaches only 4.7% of the peak, so it is shown here magnified 20×. The strip above shows the pattern as it appears on a screen, with faint bands brightened so they reproduce. Generated with matplotlib; see `scripts/figures/`.

The Secondary Maxima

Where are the maxima between the zeros? Setting $\mathrm{d}I/\mathrm{d}\beta = 0$ gives the transcendental condition $\tan(\beta/2) = \beta/2$, which has no closed-form solution. Solving numerically puts the secondary maxima at

$$\frac{\beta}{2} = 1.4303\pi,\ 2.4590\pi,\ 3.4709\pi,\ \ldots \tag{5.11}$$

slightly *inside* the naive halfway positions 1.5π, 2.5π, 3.5π — pulled toward the center because the envelope is falling. Relative to the central peak, the first three secondary maxima reach only

$$4.7\%, \qquad 1.6\%, \qquad 0.83\% \tag{5.12}$$

of I_0, falling off roughly as $1/m^2$. This is why a single-slit pattern does not look like a row of comparable fringes at all: it looks like one dominant bright band, flanked by faint ripples that most observers overlook entirely. It is also why Figure 5.5 has to magnify them 20× before they are visible on the page.

Double-Slit Diffraction: Both Effects Together

Now put the two chapters together. A *real* double slit consists of two openings, each of finite width a, whose centers are separated by $d > a$. Both effects operate at once, and — because the field from each slit is the single-slit field, and the two slits then interfere — the intensities simply multiply:

$$I(\theta) = I_0 \underbrace{\cos^2\!\left(\frac{\pi d \sin\theta}{\lambda}\right)}_{\text{interference, set by } d} \underbrace{\left[\frac{\sin(\beta/2)}{\beta/2}\right]^2}_{\text{diffraction envelope, set by } a}, \qquad \beta = \frac{2\pi}{\lambda} a \sin\theta. \tag{5.13}$$

Since $d > a$, the $\cos^2$ factor oscillates rapidly and the sinc^2 factor varies slowly: fine fringes under a broad envelope, as in Figure 5.6.

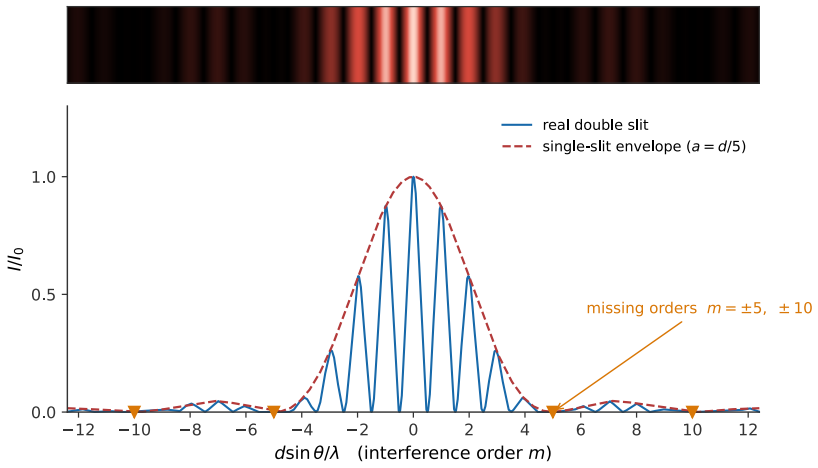

Figure 5.6: A real double slit with $d = 5a$. The rapid $\cos^2$ fringes are set by the slit separation d; the dashed envelope is the single-slit sinc^2 pattern set by the slit width a. Orders $m = \pm 5$ and ± 10 vanish because they coincide with zeros of the envelope. Generated with matplotlib; see `scripts/figures/`.

Missing Orders

The striking feature of Figure 5.6 is that some interference maxima are simply absent. An interference maximum at $d\sin\theta = m\lambda$ is wiped out if a diffraction minimum, $a\sin\theta = m'\lambda$, falls at the same angle. Dividing one condition by the other, this happens when

$$\frac{m}{m'} = \frac{d}{a}, \tag{5.14}$$

that is, whenever $m\,(a/d)$ is an integer. If d/a is itself an integer, every (d/a)-th order is missing — for $d = 5a$ the missing orders are $m = \pm 5, \pm 10, \pm 15, \ldots$, exactly as in the figure. If d/a is a ratio of small integers, say $d/a = 5/2$, orders that are multiples of 5 still vanish but nothing else does. If d/a is irrational, no order is exactly missing, though those near an envelope zero are suppressed almost to nothing.

Missing orders are a measurement, not a curiosity. Counting how many interference fringes fit inside the central diffraction maximum can reveal d/a, without measuring either d or a: the central envelope spans $|\sin\theta| < \lambda/a$ and fringes are spaced by λ/d in $\sin\theta$. When $d/a = q$ is an integer, the visible orders are $|m| \le q - 1$, giving $2q - 1$ bright fringes inside the central envelope; the orders at $|m| = q$ are missing at its edges. For a noninteger ratio, use the integer orders satisfying $|m| < d/a$ instead; the count then gives the corresponding integer bound, while the missing-order pattern can identify a rational ratio. Looking at a laboratory double-slit pattern and counting bright fringes between the first envelope zeros is therefore a genuine measurement of a ratio of two dimensions far too small to measure with a ruler.

Worked Example: Reading a Double-Slit Pattern

A double slit has slit width $a = 0.0200$ mm and separation $d = 0.100$ mm, illuminated at $\lambda = 633$ nm. (a) Which orders are missing? (b) How many bright fringes lie inside the central diffraction maximum? (c) Find the angular width of the central envelope.

(a) $d/a = (0.100\text{ mm})/(0.0200\text{ mm}) = 5$, an integer, so orders $m = \pm5, \pm10, \pm15, \ldots$ are missing.

(b) The central envelope runs from $\sin\theta = -\lambda/a$ to $+\lambda/a$; the fringes at its edges are the missing $m = \pm5$. The visible orders inside are $m = -4, \ldots, +4$: **nine bright fringes**, of which the central one is brightest.

(c) $\sin\theta_{\text{env}} = \lambda/a = (633 \times 10^{-9}\text{ m})/(2.00 \times 10^{-5}\text{ m}) = 0.0317$, so the envelope's first zeros are at $\theta = \pm1.81°$, giving a full angular width of 3.63°. On a screen 2.00 m away that is a central band 12.7 cm wide, containing nine fringes spaced $\lambda L/d = 12.7$ mm apart — a comfortable laboratory pattern.

5.3 Diffraction Gratings

A **diffraction grating** is the N-slit device of Chapter 4 built for real: thousands of parallel slits per millimeter, ruled onto glass (a transmission grating) or onto a reflective surface (a reflection grating, which is what most modern spectrometers use). A compact disc is an accidental reflection grating, its data tracks spaced 1.6 μm apart; a DVD's tracks are 0.74 μm apart, which is why it throws colors more widely.

Everything derived for N ideal slits carries over. The principal maxima satisfy the **grating equation**

$$d\sin\theta = m\lambda, \qquad m = 0, \pm1, \pm2, \ldots \tag{5.15}$$

with d the spacing between adjacent rulings; their width scales as $1/N$ and their peak intensity as N^2. What is new is only that N is now enormous — a 1 cm beam on a grating with 600 lines per millimeter illuminates 6000 lines at once — so the maxima are not fringes but needle-sharp spectral lines. And of course the finite width of each opening imposes a diffraction envelope, exactly as for the double slit, so some orders are weak or missing. (Precision gratings are *blazed*: each groove is cut at an angle chosen to steer the envelope's peak into a chosen order, so that most of the light ends up where it is wanted rather than in the useless $m = 0$ direction.)

Dispersion

Because the diffraction angle depends on λ, a grating spreads white light into a spectrum — one spectrum per order, on each side of the center. Differentiating the grating equation gives the **angular dispersion**

$$\frac{\mathrm{d}\theta}{\mathrm{d}\lambda} = \frac{m}{d\cos\theta}, \tag{5.16}$$

so the spectrum spreads out more in higher orders and with finer rulings. It also shows that the orders eventually overlap: the red end of order m can fall on top of the blue end of order $m+1$, which is why spectrometers often place a coarse filter in front of the detector to isolate a single order.

Since $|\sin\theta| \le 1$, the mathematical upper bound on the order is $m_{\max} = \lfloor d/\lambda \rfloor$ — a grating so fine that $d < \lambda$ produces no spectrum at all beyond $m = 0$. If d/λ is exactly an integer, that limiting order is at grazing angle ($\theta = 90°$) and is not usable in an ordinary detector geometry.

Tip

Before trusting any computed grating or Bragg angle, check that the sine you solved for is actually ≤ 1. Both $d\sin\theta = m\lambda$ and $n\lambda = 2d\sin\theta$ are linear in the order and carry no built-in ceiling, so plugging in one order too many silently returns a mathematically valid but physically nonexistent angle. Get in the habit of computing $m_{\max} = \lfloor d/\lambda \rfloor$ (or $n_{\max} = \lfloor 2d/\lambda \rfloor$ for Bragg reflection, later in this chapter) *before* solving for individual angles, so you already know which orders exist.

Resolving Power

The property that makes gratings scientific instruments rather than ornaments is their **resolving power**: the ability to display two nearly equal wavelengths as two separate lines rather than one blur.

Adopt the Rayleigh criterion — two lines are *just resolved* when the principal maximum of one falls on the first zero of the other. The principal maximum for wavelength λ in order m sits at $d\sin\theta = m\lambda$, and its first zero is $\lambda/(Nd\cos\theta)$ away in angle. The line at $\lambda + \Delta\lambda$ sits $m\,\Delta\lambda/(d\cos\theta)$ away. Setting these equal,

$$\frac{m\,\Delta\lambda}{d\cos\theta} = \frac{\lambda}{Nd\cos\theta} \qquad \Longrightarrow \qquad \boxed{R \equiv \frac{\lambda}{\Delta\lambda} = mN} \tag{5.17}$$

Figure 5.7 shows the criterion in action. The resolving power depends only on the order and on the number of lines actually illuminated. Two consequences are worth noting. First, resolving power improves with the *illuminated* width of the grating, not with the total number of lines ruled on it; underfilling a grating with a narrow beam throws resolution away. Second, working in second or third order doubles or triples the resolution, at the cost of dimmer lines and increased risk of overlapping orders — a trade every spectroscopist makes.

Worked Example: Resolving the Sodium Doublet

The **sodium doublet** — the pair of yellow lines responsible for the color of a sodium street lamp — consists of lines at $\lambda_1 = 589.0$ nm and $\lambda_2 = 589.6$ nm. How many grating lines must be illuminated to resolve them in first order?

The required resolving power is

$$R = \frac{\lambda}{\Delta\lambda} = \frac{589.0\ \text{nm}}{0.6\ \text{nm}} = 982. \tag{5.18}$$

From $R = mN$ with $m = 1$,

$$N = \frac{R}{m} = 982\ \text{lines}. \tag{5.19}$$

On a grating with 600 lines per millimeter this needs a beam only 1.64 mm wide — trivially met by any laboratory setup, which is why the sodium doublet is a standard first-week demonstration. In second order only 491 lines would be needed. By contrast, resolving the hyperfine structure *within* one of those lines requires a much larger resolving power than resolving the doublet — far beyond this simple demonstration and typically measured with specialized spectroscopic methods instead.

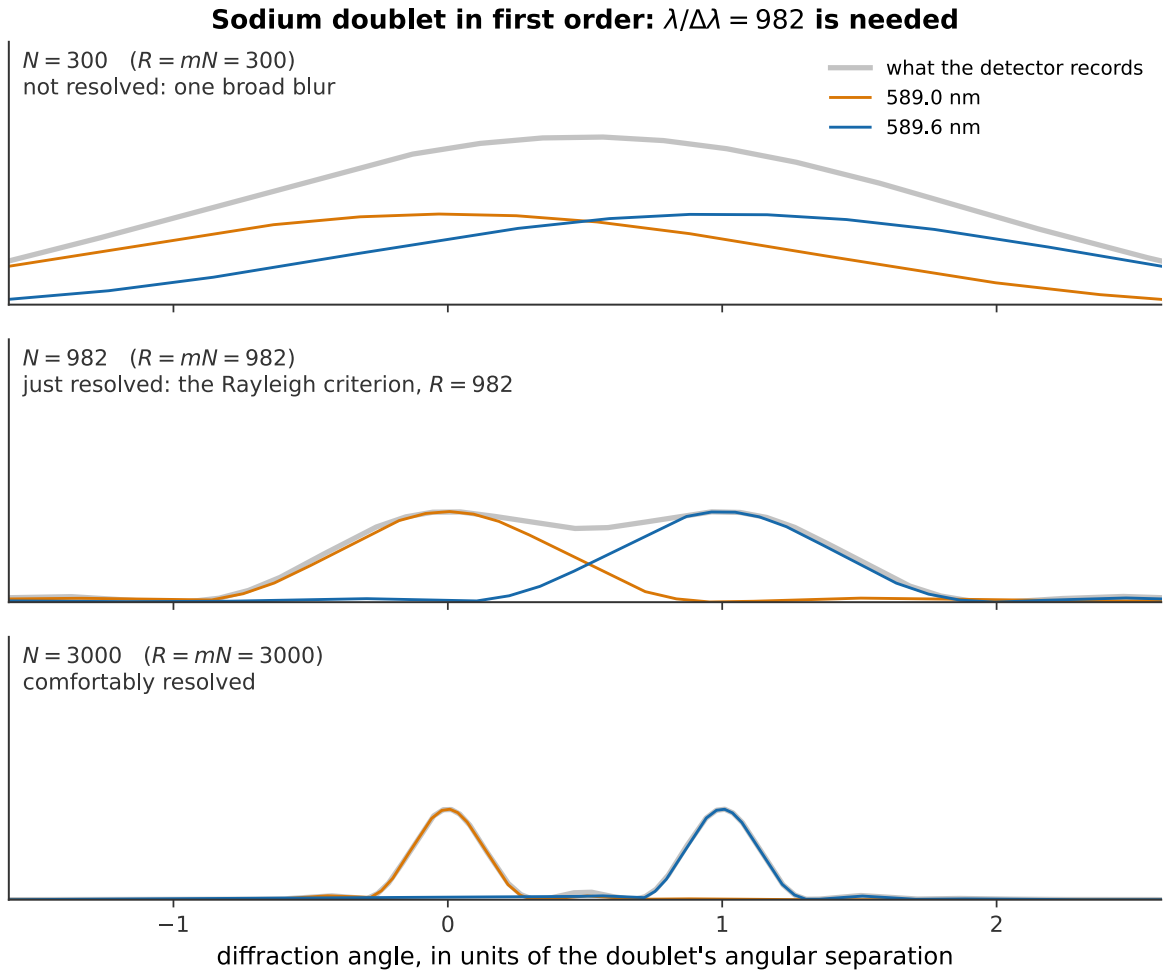

Figure 5.7: Resolving the sodium doublet (589.0 and 589.6 nm) in first order. With too few illuminated lines the two principal maxima are broader than their separation and merge into one feature. At $N = 982$, $R = mN$ equals $\lambda/\Delta\lambda$ and the peak of one line falls on the first zero of the other — the Rayleigh criterion. With $N = 3000$ the lines are cleanly separated. Generated with matplotlib; see `scripts/figures/`.

Worked Example: A Grating Spectrum

A transmission grating has 5000 lines per centimeter. (a) Find the line spacing. (b) Find the angle of the first-order maximum for $\lambda = 500$ nm. (c) Find the highest observable order. (d) Find the angular separation between the 486 nm and 656 nm hydrogen lines in first order.

(a) $d = (1\text{ cm})/5000 = 2.00 \times 10^{-6}$ m $= 2.00\ \mu$m.

(b) $\sin\theta_1 = \lambda/d = (500 \times 10^{-9}\text{ m})/(2.00 \times 10^{-6}\text{ m}) = 0.250$, so $\theta_1 = 14.5°$.

(c) $m_{\max} = \lfloor d/\lambda \rfloor = \lfloor 4.00 \rfloor = 4$, but $m = 4$ gives $\sin\theta = 1.00$ exactly — grazing emergence, unobservable in practice. The highest usable order is $m = 3$, at $\theta = 48.6°$.

(d) $\theta(486\text{ nm}) = \arcsin(0.243) = 14.06°$ and $\theta(656\text{ nm}) = \arcsin(0.328) = 19.14°$, a separation of $5.08°$. A grating spreads the visible spectrum over several degrees in first order — enough to project a spectrum across a wall.

5.4 Resolution, Bragg Diffraction, and Holography

Circular Apertures and the Limits of Resolution

Every real optical instrument — a camera lens, a telescope mirror, a microscope objective, the pupil of the eye — collects light through a *circular* aperture of diameter D. Its diffraction pattern is the circular analog of the single-slit pattern: a bright central disk, the **Airy disk**, surrounded by faint concentric rings.

The mathematics is harder than the slit case, because the integral runs over a disk rather than a line, and the answer involves a Bessel function rather than a sine. The structure of the result, however, is identical, differing only by a numerical factor that comes from the circular geometry: the first dark ring lies at

$$\sin\theta_{\min} = 1.22\,\frac{\lambda}{D} \approx \theta_{\min}. \tag{5.20}$$

Compare this with λ/a for a slit. Nothing has changed except the 1.22 — which is, in fact, the first zero of the Bessel function J_1 divided by π.

Where the factor 1.22 comes from

For a circular aperture of diameter D, the far-field amplitude at angle θ is proportional to the two-dimensional analog of the single-slit integral, evaluated over a disk instead of a line:

$$E(\theta) \propto \frac{2J_1(kR\sin\theta)}{kR\sin\theta}, \qquad k = \frac{2\pi}{\lambda}, \quad R = \frac{D}{2}, \tag{5.21}$$

where J_1 is the Bessel function of the first kind, order one — the circular-geometry replacement for the $\sin x/x$ that came from integrating across a slit. This combination plays exactly the role $\operatorname{sinc}(\beta/2)$ played in the single-slit case: it equals 1 at zero argument, giving the central Airy peak, and has its first zero at the first zero of J_1 itself,

$$kR\sin\theta_{\min} = 3.8317\ldots \tag{5.22}$$

Substituting $k = 2\pi/\lambda$ and $R = D/2$ and solving for $\sin\theta_{\min}$,

$$\sin\theta_{\min} = \frac{3.8317}{\pi}\,\frac{\lambda}{D} = 1.2197\,\frac{\lambda}{D} \approx 1.22\,\frac{\lambda}{D}. \tag{5.23}$$

The extra factor, compared with a slit's plain λ/D, is a purely geometric consequence of integrating over a disk rather than a line. No new physics enters — only a different shape of aperture.

The Rayleigh Criterion

Two distant point sources — two stars, or two features on a microscope slide — each produce their own Airy pattern in the image. If the sources are close together in angle, the two patterns overlap and merge into a single blob. Lord Rayleigh proposed a workable convention for where to draw the line: two sources are **just resolved** when the center of one Airy disk falls on the first dark ring of the other. That is, when their angular separation equals

$$\theta_{\min} = 1.22\,\frac{\lambda}{D} \qquad (\textbf{Rayleigh criterion}). \tag{5.24}$$

Figure 5.8 shows what this looks like. At the criterion there is a shallow dip between the two peaks — about 26% below the maxima. The criterion is a convention, not a law of nature: with a high signal-to-noise ratio and a known point-spread function, computational methods can do better, and some super-resolution fluorescence methods exploit temporal control or sparse emitters to distinguish features closer together than the Rayleigh separation. But as an estimate of what an instrument can do, it is excellent.

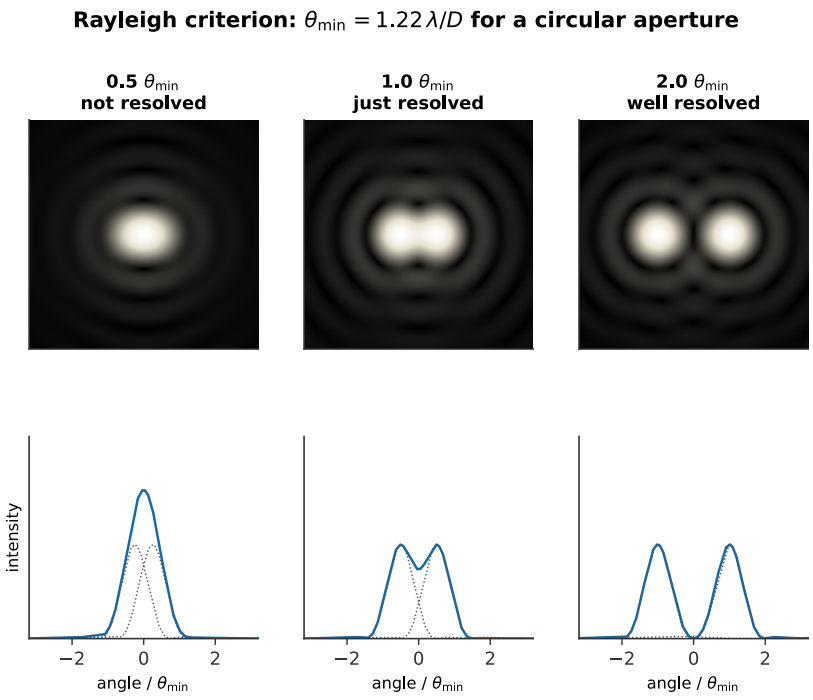

Figure 5.8: Two point sources imaged through a circular aperture, at separations of 0.5, 1.0, and 2.0 times $\theta_{\min} = 1.22\lambda/D$, with intensity profiles beneath. At half the Rayleigh separation the pair is indistinguishable from a single elongated source; at the Rayleigh separation a shallow dip appears between the peaks; at twice it the two sources are unmistakable. Generated with matplotlib; see `scripts/figures/`.

The crucial thing about the Rayleigh criterion is its universality. It applies to every instrument with aperture D, however perfectly made. It is not an engineering shortcoming to be designed away; it is a consequence of light having a wavelength and the instrument having an edge. The only levers available are λ and D: use a shorter wavelength, or build a bigger aperture.

Both levers are pulled hard in practice. Telescopes get larger, and radio astronomers

link dishes thousands of kilometers apart to synthesize an aperture the size of the Earth. Photolithography, which prints circuit features by projecting a mask onto a silicon wafer, has driven its light source from visible wavelengths down through deep ultraviolet to today's extreme-ultraviolet at 13.5 nm for exactly this reason. And electron microscopy takes the argument to its conclusion by abandoning light altogether in favor of electron matter waves whose wavelength (Chapter 7) can be a hundred thousand times shorter than visible light.

Worked Example: The Hubble Space Telescope

The Hubble Space Telescope has a primary mirror of diameter $D = 2.4$ m. (a) Estimate its diffraction-limited resolution at $\lambda = 550$ nm. (b) What is the smallest feature it could resolve on the Moon, 3.84×10^8 m away?

(a)

$$\theta_{\min} = 1.22\,\frac{\lambda}{D} = 1.22\,\frac{550 \times 10^{-9}\ \text{m}}{2.4\ \text{m}} = 2.80 \times 10^{-7}\ \text{rad} = 0.058\ \text{arcsec}. \tag{5.25}$$

(b)

$$s = \theta_{\min}\, d = (2.80 \times 10^{-7}\ \text{rad})(3.84 \times 10^{8}\ \text{m}) = 107\ \text{m}. \tag{5.26}$$

Hubble could not photograph the Apollo landing sites; the largest hardware left there is a few meters across. The limitation is not the telescope's quality but its aperture, and no amount of image processing changes the fact that the information never entered the instrument.

Worked Example: The Eye, and Why Ground Telescopes Are Different

(a) The pupil of the eye is about $D = 4.0$ mm in ordinary lighting. Its diffraction limit at $\lambda = 550$ nm is

$$\theta_{\min} = 1.22\,\frac{550 \times 10^{-9}\ \text{m}}{4.0 \times 10^{-3}\ \text{m}} = 1.68 \times 10^{-4}\ \text{rad} \approx 0.6\ \text{arcmin}. \tag{5.27}$$

At a reading distance of 25 cm this corresponds to $s = (1.68 \times 10^{-4}\ \text{rad})(0.25\ \text{m}) = 42\ \mu\text{m}$ — about half the width of a human hair, and in reasonable agreement with measured visual acuity of about 1 arcmin. The eye is, remarkably, working close to its diffraction limit.

(b) Now apply the same formula to a 2.4 m telescope on the ground: $\theta_{\min} = 0.058$ arcsec, the same as Hubble. Yet ground-based images at visible wavelengths are typically blurred to about 1 arcsec, a factor of 17 worse. The culprit is not diffraction but **atmospheric seeing**: turbulent cells of air, of order 10 cm across, refract the incoming wavefront by slightly different amounts and scramble it. The effective aperture is set by the turbulence, not the mirror. Two responses are possible — put the telescope above the atmosphere, which is why Hubble exists, or measure the distortion hundreds of times a second and cancel it with a deformable mirror, which is *adaptive optics*, and which now lets large ground telescopes reach their diffraction limits in the infrared.

This example is worth keeping in mind whenever a diffraction limit is quoted: it is a ceiling on performance, not a promise of it.

X-Ray Diffraction and Bragg's Law

Gratings work because their line spacing is comparable to the wavelength of the light. To probe the arrangement of atoms in a solid, spaced a few tenths of a nanometer apart, we would need a grating with that spacing — and nature supplies one, in the form of a crystal, provided we bring a wave with a matching wavelength. That means **X-rays**, with $\lambda \sim 0.1$ nm.

> Crystallographers often quote spacings in **angstroms** (1 Å $\equiv 10^{-10}$ m $= 0.1$ nm), a unit sized conveniently to typical interatomic distances. The copper K_α wavelength of 0.154 nm used in the worked example below is equivalently 1.54 Å.

Max von Laue proposed the experiment in 1912 and his collaborators Friedrich and Knipping performed it, obtaining a pattern of discrete spots from a copper sulfate crystal that established two things at once: that X-rays are waves, and that crystals are periodic arrays of atoms. Both had been conjectures until that afternoon.

W. H. Bragg and his son W. L. Bragg supplied the simple picture that made the technique usable, shown in Figure 5.9. Regard the crystal as a stack of parallel atomic planes separated by d, and consider X-rays reflecting specularly from successive planes. By long-standing crystallographic convention, the angle θ is measured *from the plane itself*, not from the normal — a trap for anyone importing habits from optics.

> **Warning**
> The Bragg angle θ is measured from the reflecting *plane*, not from the normal to it — the opposite of the convention used for mirrors and lenses elsewhere in optics. A beam that grazes the planes at nearly 0° corresponds to *small* θ in Bragg's law, not large. Carrying over the normal-incidence habit from geometric optics is the single most common source of a sine-versus-cosine error in Bragg's law calculations.

Ray 2 penetrates one layer deeper. Measured between the wavefronts through the point A where ray 1 reflects, it travels two extra legs — one on the way in, one on the way out — each of length $d\sin\theta$. Constructive interference between rays reflected from successive planes therefore requires

$$n\lambda = 2d\sin\theta \qquad (\textbf{Bragg's law}), \qquad n = 1, 2, 3, \ldots \tag{5.28}$$

The consequence is that a crystal reflects a given X-ray wavelength strongly *only* at particular angles. Rotate a crystal in a monochromatic X-ray beam and the detector sees nothing at all until an angle satisfying Bragg's law comes up, at which point a sharp peak appears. Measuring the set of such angles, for a set of crystal orientations, determines the spacings and symmetry of the atomic planes.

This is **X-ray crystallography**, and its record is hard to overstate: the structures of table salt and diamond (the Braggs, 1913), of penicillin and vitamin B_{12} (Dorothy Hodgkin), of DNA — the double helix inferred by Watson and Crick from the diffraction photographs taken by Rosalind Franklin and Raymond Gosling — and of hundreds of thousands of proteins since. Nearly everything we know about where atoms sit in condensed matter came through Bragg's law.

The lattice on the other side of Bragg's law is the subject of Figure 5.10. Changing the unit cell changes the set of atomic planes, their spacings d, and therefore the angles at which

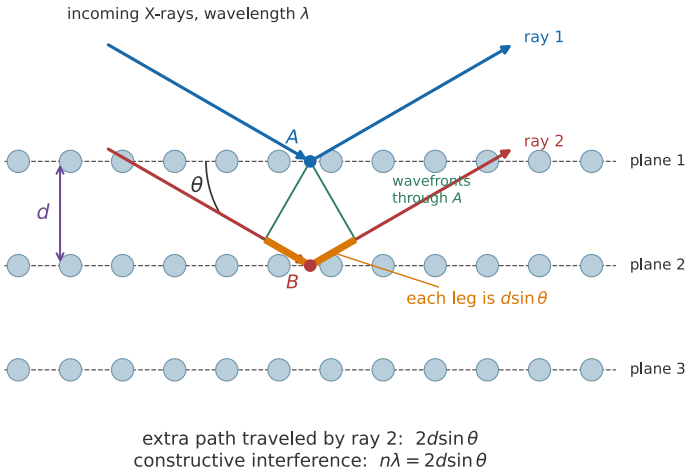

Figure 5.9: Bragg reflection. Two parallel X-rays strike successive atomic planes separated by d. Relative to the wavefronts through A, the ray reflecting from the lower plane travels two extra legs, each of length $d\sin\theta$. Constructive interference requires the total, $2d\sin\theta$, to be a whole number of wavelengths. Original schematic generated with matplotlib; see `scripts/figures/`.

reflections appear — including the aperiodic case, where the diffraction pattern is as sharp as a crystal's but carries a symmetry no crystal can have.

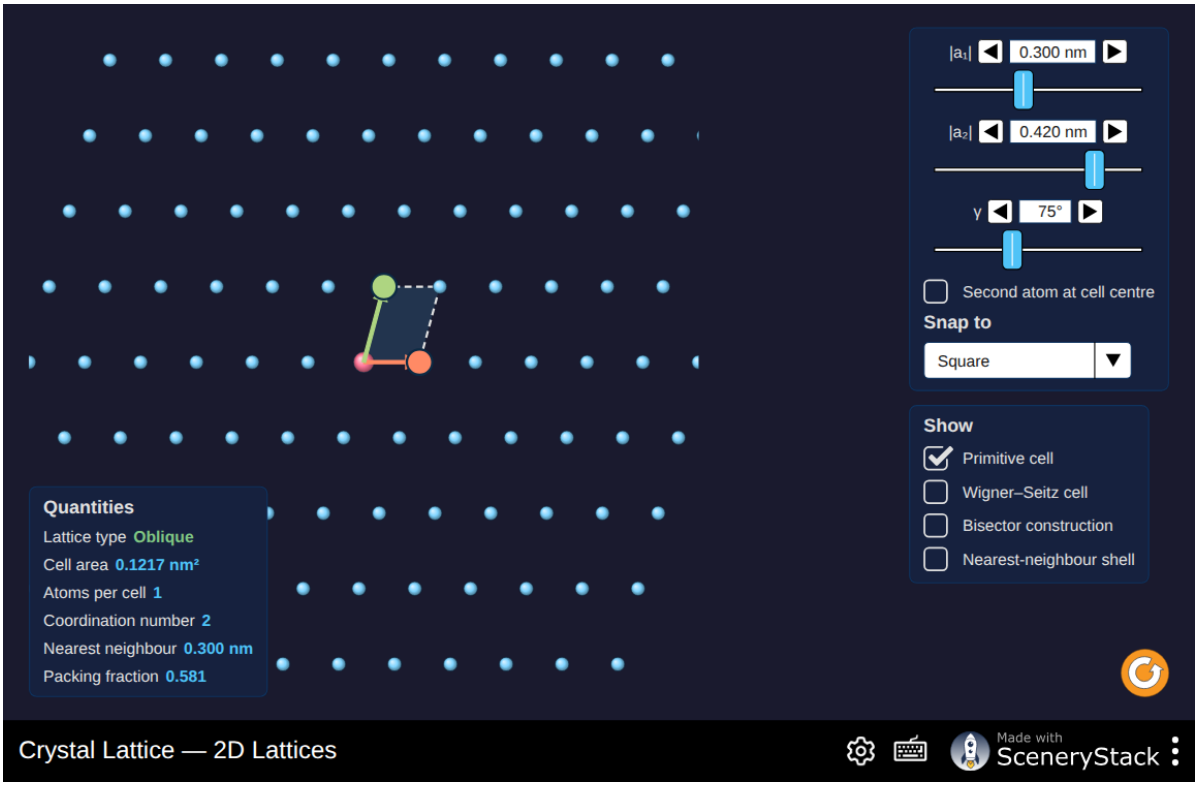

Figure 5.10: Crystal structure in five screens: the two-dimensional Bravais lattices, cubic unit cells, close packing, Miller indices, and aperiodic order. The Miller-index screen is the direct companion to Bragg's law — it is where the planes of spacing d come from. Interactive simulation: Crystal Lattice

Worked Example: An Interplanar Spacing

X-rays of wavelength $\lambda = 0.154$ nm (the copper K_α line, the workhorse of laboratory diffractometers) produce a first-order Bragg reflection at $\theta = 22.5°$. Find the spacing of the reflecting planes, and find where the second-order reflection appears.

First order, $n = 1$:

$$d = \frac{n\lambda}{2\sin\theta} = \frac{0.154\text{ nm}}{2\sin 22.5°} = \frac{0.154\text{ nm}}{2(0.3827)} = 0.201\text{ nm}, \tag{5.29}$$

a typical interatomic spacing. For $n = 2$,

$$\sin\theta_2 = \frac{2\lambda}{2d} = \frac{0.154\text{ nm}}{0.201\text{ nm}} = 0.766 \quad\Longrightarrow\quad \theta_2 = 50.0°. \tag{5.30}$$

Note that a third order would require $\sin\theta_3 = 1.149 > 1$: with this wavelength and this spacing, only two orders exist. Requiring $n\lambda \le 2d$ also shows why X-rays are essential — with visible light, $\lambda \gg 2d$ and *no* order exists at all, which is exactly why crystals look like ordinary transparent solids rather than gratings.

Not Just X-Rays

Bragg's law contains no reference to what kind of wave is diffracting. Any wave whose wavelength is comparable to the interplanar spacing will do — and by the time you reach Chapter 7, you will know that electrons and neutrons are such waves.

That is not a rhetorical flourish; it is the historical route by which matter waves were confirmed. In 1927 Clinton Davisson and Lester Germer fired low-energy electrons at a nickel crystal and found the scattered intensity peaking at exactly the angles Bragg's law predicts, for a wavelength matching de Broglie's $\lambda = h/p$. Neutron diffraction, which is sensitive to light atoms in a way X-rays are not, is now the standard way to locate hydrogen in a crystal structure and to map magnetic ordering. The same equation, first written for X-rays, ended up as one of the sharpest pieces of evidence for the wave nature of matter.

Holography

An ordinary photograph records intensity and throws away phase, which is why it is flat: all the information about the *shape* of the wavefront arriving from the scene is lost. A **hologram**, invented by Dennis Gabor in 1947, records the phase as well — by the only means available, namely by converting phase differences into intensity differences through interference.

The recording step splits a laser beam in two. One part, the *object beam*, illuminates the subject and scatters from it onto the film. The other, the *reference beam*, goes straight to the film. Where the two arrive in phase the film is exposed; where they arrive out of phase it is not. What the film records is therefore a dense, fine interference pattern — nothing resembling a picture, just an apparently random speckle of microscopic fringes — which encodes both the amplitude and the phase of the light arriving from every point of the object.

The reconstruction step is pure diffraction. Illuminate the developed film with the reference beam alone, and the recorded fringe pattern acts as a complicated diffraction grating, diffracting the reference beam into exactly the wavefront that the object originally sent out. An observer looking into the film sees that wavefront and cannot tell it from the object: it has full parallax, and looking around the edge of a foreground object reveals what is behind it.

Two properties follow directly from this account. First, holography demands a source with a coherence length longer than the depth of the scene, which is why it was impossible before the laser and why Gabor's original demonstrations were so limited. Second, every part of the hologram receives light from every part of the object, so a hologram cut in half still reconstructs the whole scene — from a smaller effective aperture, and therefore, by the Rayleigh criterion, with correspondingly poorer resolution.



Figure 5.11: A hands-on look at diffraction gratings and embossed rainbow holograms outside the lab, using the same dispersion behavior developed earlier in this chapter. Watch on YouTube: Using Quantum Mechanics to Make Holographic Rainbows on Chocolate

Looking Ahead: Diffraction and the Uncertainty Principle

There is one more thing to extract from single-slit diffraction, and it is the reason this chapter belongs in a book on modern physics.

Take the pattern seriously as a statement about photons. A photon that passes through a slit of width a has, at that moment, a position along the slit known to within

$$\Delta y \approx a. \tag{5.31}$$

It arrives traveling straight ahead, with no transverse momentum. Yet it lands somewhere within the central diffraction maximum, which means that on emerging it has acquired a

transverse momentum of order $p \sin\theta_1$, where θ_1 is the angle to the first minimum. Using $\sin\theta_1 = \lambda/a$ and the photon momentum $p = h/\lambda$ (established in Chapter 6),

$$\Delta p_y \approx p \sin\theta_1 = \frac{h}{\lambda} \cdot \frac{\lambda}{a} = \frac{h}{a}. \tag{5.32}$$

Multiply:

$$\Delta y \, \Delta p_y \approx a \cdot \frac{h}{a} = h. \tag{5.33}$$

The width of the slit has canceled out. Squeeze the slit to pin down the photon's position and the diffraction pattern widens by exactly enough to keep the product fixed; open it up and the reverse happens. There is no way to make both small at once — not because the measuring apparatus is clumsy, but because the light is a wave, and this relation is a property of waves.

This is the **Heisenberg uncertainty principle**, arrived at from nothing but classical wave optics and one quantum fact ($p = h/\lambda$). Chapter 7 derives it properly, shows that it applies to electrons and every other particle, and works out its consequences. But it is worth seeing now, in this chapter, that its origin lies in the ordinary mathematics of diffraction — that the strangest feature of quantum mechanics is, at bottom, the same phenomenon that spreads light behind a narrow slit.

5.5 Summary

- Diffraction and interference are the same physics — coherent waves added as fields. The words distinguish a continuum of sources (an integral) from a discrete set (a sum), and nothing more.
- The **Huygens–Fresnel principle** treats each point of an aperture as a source of wavelets that are summed *with their phases*. **Fraunhofer** (far-field) diffraction applies when the Fresnel number $F = a^2/L\lambda \ll 1$, or whenever a lens is used to bring parallel bundles to a focus; **Fresnel** (near-field) diffraction applies otherwise, and produces effects such as the Poisson spot.
- **Single-slit minima** occur at $a \sin\theta = m\lambda$ for nonzero integer m, derived by pairing points $a/2m$ apart so that every pair cancels. $m = 0$ is excluded: the center is a broad maximum.
- The **single-slit intensity** is $I = I_0[\sin(\beta/2)/(\beta/2)]^2$ with $\beta = (2\pi/\lambda)a\sin\theta$, obtained from the chord of the circular phasor arc. The central maximum has half-width $\theta_1 \approx \lambda/a$, is twice as wide as the side lobes, and dominates: the first side lobe reaches only 4.7% of the peak.
- **Narrower slits spread light more.** For $a \leq \lambda$ no minimum exists and the slit radiates like a point source, justifying the idealization of Chapter 4.
- A **real double slit** gives the product $I = I_0 \cos^2(\pi d \sin\theta/\lambda)[\sin(\beta/2)/(\beta/2)]^2$: fine interference fringes set by d, under a broad envelope set by a. Orders for which $m\,a/d$ is an integer are **missing**; counting the fringes inside the central envelope measures d/a.
- A **diffraction grating** obeys $d\sin\theta = m\lambda$ with angular dispersion $\mathrm{d}\theta/\mathrm{d}\lambda = m/(d\cos\theta)$ and resolving power $R = \lambda/\Delta\lambda = mN$, where N is the number of *illuminated* lines. Only orders up to $m_{\max} = \lfloor d/\lambda \rfloor$ exist.
- A **circular aperture** of diameter D produces an Airy pattern with first zero at $\theta_{\min} = 1.22\lambda/D$. The **Rayleigh criterion** adopts this as the resolution limit of any instrument

with that aperture — a fundamental diffraction limit, though real performance may be worse (atmospheric seeing) or, with enough signal and cleverness, better.

- **Bragg's law**, $n\lambda = 2d\sin\theta$ with θ measured from the atomic planes, governs constructive reflection from a crystal lattice. It underlies X-ray crystallography, and applies equally to electron and neutron waves — as Davisson and Germer showed in 1927 (Chapter 7).
- A **hologram** records the interference between an object beam and a reference beam, capturing phase as well as amplitude; illuminating it with the reference beam diffracts a reconstruction of the original wavefront.
- Applying $\Delta y \approx a$ and $\Delta p_y \approx h/a$ to a single slit gives $\Delta y\,\Delta p_y \approx h$: the **uncertainty principle**, already visible in classical diffraction.

5.6 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens on a single-slit diffraction pattern.

1. **Multiple choice, from a figure.** What change makes a single-slit diffraction pattern wider? Figure description: Single-slit diffraction intensity pattern with a broad central maximum and weaker side lobes. Choices: (A) Narrowing the slit. (B) Widening the slit. (C) Increasing the screen distance while measuring angles. (D) Adding more illuminated grating lines without changing the slit.

2. **True or false.** The value m = 0 is a single-slit diffraction minimum.

3. **Drag the words.** Complete the principal diffraction relations. Single-slit minima obey _____. A grating's resolving power is _____. A circular aperture has Rayleigh angle _____. Crystal planes satisfy _____. (Terms: a sin $\theta = m\lambda$; R = mN; $1.22\lambda/D$; $n\lambda$ = 2d sin θ; d sin $\theta = m\lambda$; λ/D^2.)

4. **Fill in the blanks.** Complete the grating and aperture results. A grating used in order m = 2 with N = 500 illuminated lines has resolving power _____, while doubling a circular aperture's diameter makes its diffraction-limited angle _____ as large.

5. **Mark the words.** Mark the two changes that improve diffraction-limited resolution. "A grating resolves closer wavelengths by using more illuminated lines. A circular instrument resolves closer angular sources by using a larger aperture diameter. Narrowing an ordinary slit broadens its pattern."

5.7 Conceptual Questions

1. State, in one sentence each, the physical difference between the conditions $d\sin\theta = m\lambda$ for a double slit and $a\sin\theta = m\lambda$ for a single slit. Why do two nearly identical formulas describe opposite things?

2. Explain, using the pairing argument, why $\theta = 0$ can never be a diffraction minimum, no matter how many strips the slit is divided into.

3. A single slit is illuminated with red light, then with blue light. Which pattern is wider, and why? Now the same slit is made narrower. Which way does the pattern change? Are the two answers consistent with a single rule?

4. Explain why a wider telescope aperture *improves* angular resolution, while a wider slit *narrows* the diffraction pattern. These sound like opposite statements about what a bigger aperture does. Show that they are the same statement.

5. In a real double-slit pattern, some interference maxima are missing entirely. Explain what has happened to the light that "should" have been there, and why energy is nonetheless conserved.

6. Why can visible light not be used to determine crystal structures by Bragg diffraction? Answer quantitatively, using the constraint $n\lambda \le 2d$.

7. A hologram is cut in half. Explain why each half still reconstructs the entire scene, and what is degraded.

8. A camera lens is stopped down from $f/2$ to $f/16$, reducing the aperture diameter by a factor of 8. Aberrations improve when a lens is stopped down, but diffraction gets worse. Explain the trade-off, and state which effect dominates at very small apertures.

5.8 Problems

Exercise 5.1
(Straightforward)

A single slit of width $a = 0.0250$ mm is illuminated at $\lambda = 633$ nm. Find the angles to the first and second diffraction minima, and the width of the central maximum on a screen 1.80 m away.

Exercise 5.2
(Straightforward)

A slit of unknown width produces a central diffraction maximum 4.00 cm wide on a screen 1.50 m away, using $\lambda = 520$ nm light. Find the slit width.

Exercise 5.3
(Straightforward)

Using $I = I_0[\sin(\beta/2)/(\beta/2)]^2$, compute the intensity, as a fraction of I_0, at the angle exactly halfway between the first and second minima ($\beta/2 = 1.5\pi$). Compare your answer with the true first secondary maximum, 4.7% at $\beta/2 = 1.4303\pi$, and comment on the size of the discrepancy.

Exercise 5.4
(Straightforward)

Show that the Fresnel number for a 0.10 mm slit at $\lambda = 550$ nm becomes of order unity at a screen distance of about 2 cm. What does this tell you about the observation distance needed for the far-field formulas of this chapter to apply?

Exercise 5.5
(Straightforward)

A double slit has slit width $a = 0.020$ mm and separation $d = 0.100$ mm, illuminated at $\lambda = 550$ nm. (a) Find d/a. (b) List the missing orders up to $m = 12$. (c) Find the number of bright fringes visible within the central diffraction maximum.

Exercise 5.6
(Straightforward)

A double slit produces 7 bright fringes inside its central diffraction maximum (counting the central fringe, and excluding the missing orders at the edges). Deduce the ratio d/a.

Exercise 5.7
(Straightforward)

A diffraction grating has 5000 lines per centimeter. (a) Find the line spacing d. (b) Find the angle of the first-order maximum for $\lambda = 500$ nm. (c) Find the highest order observable for this wavelength. (d) Find the angular dispersion $\mathrm{d}\theta/\mathrm{d}\lambda$ in first order at this wavelength, in degrees per nanometer.

Exercise 5.8
(Straightforward)

A grating with $N = 4000$ illuminated lines is used in second order. (a) Find the smallest wavelength difference it can resolve near $\lambda = 600$ nm. (b) Repeat for first order, and state the factor by which the resolution changes. (c) If the beam is narrowed so that only 1000 lines are illuminated, what happens to the resolving power, and why does the total number of lines ruled on the grating not matter?

Exercise 5.9
(Moderate)

A grating with 600 lines per millimeter is illuminated with white light (400–700 nm). (a) Find the angular width of the first-order spectrum. (b) Find the angular width of the second-order spectrum. (c) Show that the second- and third-order spectra overlap, and find the wavelength at which the overlap begins.

Exercise 5.10
(Moderate)

The human eye has a pupil diameter of about $D = 4.0$ mm in typical lighting. Using the Rayleigh criterion at $\lambda = 550$ nm, estimate (a) the smallest angular separation the eye can resolve and (b) the corresponding separation of two objects at a reading distance of 25 cm.

Exercise 5.11
(Moderate)

A car's headlights are separated by 1.3 m. Using the eye's diffraction limit from Exercise 5.10, estimate the greatest distance at which they could in principle be resolved as two separate lights. Then comment on whether atmospheric turbulence, the eye's own aberrations, or the finite spacing of retinal cells is likely to set a more restrictive limit in practice.

Exercise 5.12

(Moderate)

A telescope's objective diameter is doubled at fixed wavelength. By what factor does the smallest resolvable angle change? By what factor would it change instead if the diameter were held fixed but the observing wavelength were shifted from $\lambda = 550$ nm to $\lambda = 275$ nm?

Exercise 5.13

(Moderate)

A microscope objective of diameter 5.0 mm has a working distance of 8.0 mm. (a) Estimate the smallest separation it can resolve at $\lambda = 550$ nm, treating the objective as a circular aperture subtending the corresponding angle. (b) Estimate how much better an electron microscope would do if its electron waves have $\lambda = 4.0$ pm and the effective aperture angle is the same.

Exercise 5.14

(Moderate)

X-rays of wavelength $\lambda = 0.0709$ nm (the molybdenum K_α line) produce a first-order Bragg reflection at $\theta = 15.0^\circ$. (a) Find the interplanar spacing. (b) Find the angle of the second-order reflection. (c) Determine the highest order that exists for these planes.

Exercise 5.15

(Moderate)

A crystal has an interplanar spacing $d = 0.282$ nm, close to the value for sodium chloride. (a) Find the X-ray wavelength that produces a first-order Bragg reflection at $\theta = 20.0^\circ$. (b) Confirm that this wavelength lies in the X-ray region. (c) Find the longest wavelength that can produce any Bragg reflection at all from these planes.

Exercise 5.16

(Moderate)

Show that the general condition for a Bragg reflection to exist is $\lambda \le 2d$, and use it to explain why $\lambda = 550$ nm light produces no diffraction peaks from a crystal with $d = 0.3$ nm.

Exercise 5.17

(Challenging)

Explain physically why a narrower slit produces a wider central maximum, using $\theta_1 \approx \lambda/a$, and reconcile this with the fact that a larger telescope aperture gives finer resolution. Are the two statements in tension? Explain carefully.

Exercise 5.18

(Challenging)

A single slit of width $a = 1.0$ μm is illuminated at $\lambda = 500$ nm. (a) Find the angle to the first minimum. (b) Treating the transmitted light as photons of momentum $p = h/\lambda$, estimate Δp_y from the angular spread and verify that $\Delta y\, \Delta p_y \approx h$. (c) Repeat for $a = 0.50$ μm and confirm that the product does not change.

Exercise 5.19
(Challenging)

The Event Horizon Telescope images the black hole at the center of the galaxy M87 at a wavelength of 1.3 mm, using an array of radio dishes spread across the Earth to synthesize an aperture $D \approx 1.0 \times 10^7$ m. (a) Find its angular resolution. (b) The black hole's shadow subtends about 40 microarcseconds. Show that the array can just resolve it. (c) Explain why the same resolution cannot be achieved with a single dish.

Exercise 5.20
(Challenging)

A compact disc has data tracks spaced 1.6 μm apart and acts as a reflection grating. (a) Find the first-order diffraction angle for $\lambda = 650$ nm light at normal incidence. (b) Find the highest order that exists. (c) A DVD has tracks spaced 0.74 μm apart; explain, using your results, why a DVD spreads colors more widely than a CD.

Part III

Quantum Theory

Chapter 6

Particle Properties of Waves

Learning Objectives

By the end of this chapter, you should be able to:

- Explain the ultraviolet catastrophe and how Planck's quantization hypothesis resolves it.
- State and apply Planck's radiation law, Wien's displacement law, and the Stefan–Boltzmann law.
- Describe the photoelectric effect and explain why classical (wave) theory fails to account for its observed features.
- Apply Einstein's photon model, $E = hf$, to the photoelectric effect and interpret the stopping-potential and work-function data it predicts.
- Explain X-ray production (bremsstrahlung) as an energy-bookkeeping inverse of the photoelectric effect, and apply the Duane–Hunt relation for the minimum X-ray wavelength.
- Derive and apply the Compton scattering formula, and explain why Compton scattering demonstrates that photons carry momentum.
- Explain pair production and pair annihilation as further evidence for the particle nature of light, and use conservation of energy and momentum to find threshold conditions for each.

Introduction

Chapters 1–3 dismantled the ether and rebuilt kinematics and dynamics on Einstein's postulates, but light itself — a wave, according to Maxwell's equations, which special relativity leaves intact — remained conceptually untouched. Chapters 4–5 then developed that wave picture in full quantitative detail: interference and diffraction, both direct, precise confirmations that light obeys a wave equation. This chapter takes up a second, independent crisis of classical physics, one concerning not the *kinematics* of light but its very *nature*. Four phenomena — blackbody radiation, the photoelectric effect, X-ray production, and Compton scattering — each show, in a different experimental setting, that electromagnetic radiation exchanges energy with matter not continuously, as a wave should, but in discrete packets. The concept that emerges, the **photon**, carries energy $E = hf$ and momentum $p = h/\lambda$, and behaves in each of these experiments like a particle, even though light indisputably also shows the wave behavior of Chapters 4–5 in other settings. A final phenomenon, pair production, shows that a photon's energy can be converted entirely into matter, tying the particle nature of light directly to the mass–energy equivalence of Chapter 3. Reconciling these two faces of light is the beginning of

quantum mechanics.

6.1 Blackbody Radiation and Planck's Hypothesis

Any object at temperature T emits electromagnetic radiation across a continuous range of wavelengths, with an intensity distribution depending on T. An idealized perfect absorber and emitter is called a **blackbody**; its emitted spectral distribution, the blackbody spectrum, depends only on temperature, not on the material. A good laboratory approximation is a small hole in the wall of an otherwise closed, heated cavity: radiation entering the hole is absorbed and re-emitted many times by the cavity walls before any of it can escape, so the radiation leaking out through the hole is, to excellent approximation, in thermal equilibrium with the cavity walls and independent of what the walls are made of.

> The Boltzmann constant, $k_B = 1.381 \times 10^{-23}$ J/K, converts a temperature into a characteristic thermal energy scale $k_B T$. It recurs throughout this chapter's blackbody formulas precisely because the equipartition theorem assigns exactly this much average energy to each classical oscillator.

The Ultraviolet Catastrophe

Late-nineteenth-century classical physics attempted to compute this spectrum by treating the electromagnetic field inside the cavity as a superposition of standing-wave modes, each behaving as an independent harmonic oscillator, and invoking the equipartition theorem — every such oscillator, in thermal equilibrium at temperature T, should carry average energy $k_B T$, independent of its frequency. Counting the number of standing-wave modes per unit volume in a wavelength interval $d\lambda$ (a purely geometric problem, fixed by the cavity's size) and multiplying by $k_B T$ per mode gives the **Rayleigh–Jeans law**,

$$u(\lambda, T) = \frac{8\pi k_B T}{\lambda^4}, \tag{6.1}$$

for the spectral energy density. This matches the observed spectrum reasonably well at long wavelength, but diverges as $\lambda \to 0$: since the number of short-wavelength modes grows without bound and each is assigned the same energy $k_B T$, the predicted radiated energy — obtained by integrating $u(\lambda, T)$ over all λ — is *infinite*. This absurd prediction, sharply contradicted by the observed spectrum (which rises from zero at short wavelength, peaks, and falls off at long wavelength), became known as the **ultraviolet catastrophe**: not a small discrepancy to be patched, but a qualitative failure of classical statistical mechanics applied to the electromagnetic field.

Figure 6.1 plots both curves against each other. Raising the temperature moves the peak and lifts the total area, as the next section's displacement and Stefan–Boltzmann laws require; the classical curve, plotted alongside, runs off the top of the graph at short wavelength no matter what temperature is chosen.

Planck's Quantization Hypothesis

Max Planck resolved this in 1900 by a hypothesis with no classical justification: the oscillators making up the cavity walls cannot exchange energy with the field continuously, but only in discrete multiples of a fundamental quantum,

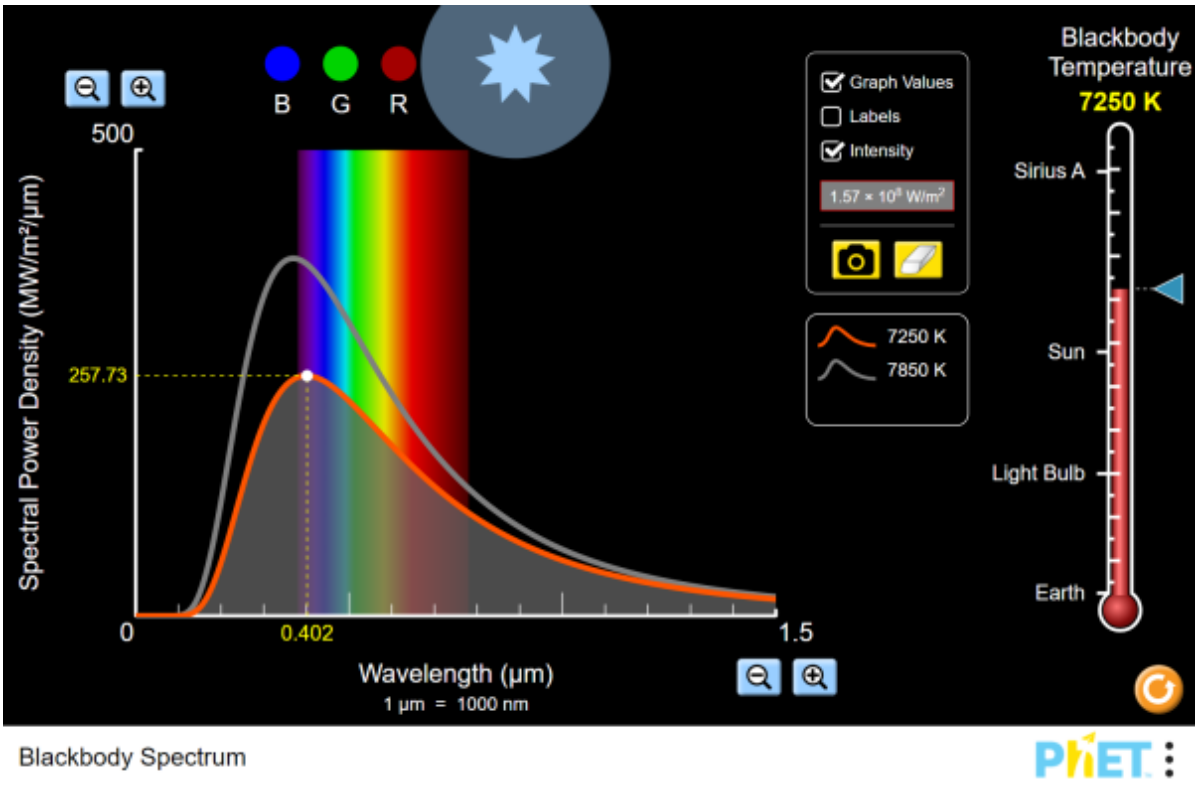

Figure 6.1: The blackbody spectrum as a function of temperature, with the Rayleigh–Jeans prediction available for comparison. The gap between the two at short wavelength is the ultraviolet catastrophe. Interactive simulation: Blackbody Spectrum

$$E_n = nhf, \qquad n = 0, 1, 2, \ldots, \tag{6.2}$$

Planck's constant is often quoted instead as $\hbar \equiv h/2\pi = 1.055 \times 10^{-34}$ J · s ("h-bar"), the form that dominates the angular-momentum and wave-mechanics formulas of later chapters. This chapter sticks to h throughout, since $E = hf$ and $p = h/\lambda$ are the natural forms for photon energy and momentum.

where f is the oscillator's frequency and h is a new fundamental constant (**Planck's constant**), $h = 6.626 \times 10^{-34}$ J·s. With this assumption, the average energy of a mode at frequency f in thermal equilibrium at temperature T is not $k_B T$ (the equipartition value, independent of f, which causes the catastrophe) but

$$\langle E \rangle = \frac{hf}{e^{hf/k_B T} - 1}, \tag{6.3}$$

which suppresses high-frequency (short-wavelength) modes because at high f, $hf \gg k_B T$ makes a single quantum too "expensive" to be thermally excited: the Boltzmann factor $e^{-hf/k_B T}$ governing how likely a mode is to hold even one quantum falls off exponentially, cutting off the ultraviolet divergence entirely. This leads to **Planck's radiation law** for the spectral energy density,

$$u(\lambda, T) = \frac{8\pi hc}{\lambda^5} \frac{1}{e^{hc/\lambda k_B T} - 1}, \tag{6.4}$$

which matches the observed blackbody spectrum at all wavelengths and temperatures, and reduces to the Rayleigh–Jeans law in the limit $hf \ll k_B T$, i.e. at long wavelength, where $e^{hf/k_B T} - 1 \approx hf/k_B T$ and quantization effects become negligible — exactly the regime in which the classical calculation had already succeeded. Planck himself regarded the quantization of oscillator energy as a mathematical device rather than a physical claim about light. It was Einstein who, five years later, proposed taking it literally.

Wien's Displacement Law and the Stefan–Boltzmann Law

Differentiating $u(\lambda, T)$ with respect to λ and setting the result to zero gives **Wien's displacement law** for the wavelength of peak emission,

$$\lambda_{\max} T = 2.898 \times 10^{-3} \text{ m} \cdot \text{K}, \tag{6.5}$$

which is why hotter objects glow at shorter (bluer) wavelengths — from the deep red of a heating element to the blue-white of a hot star. Integrating $u(\lambda, T)$ over all wavelengths gives the total power radiated per unit surface area of an ideal blackbody, the **Stefan–Boltzmann law**,

$$\frac{P}{A} = \sigma T^4, \qquad \sigma = 5.670 \times 10^{-8} \text{ W} \cdot \text{m}^{-2} \cdot \text{K}^{-4}, \tag{6.6}$$

with σ the **Stefan–Boltzmann constant**. The steep T^4 dependence means that radiated power is extremely sensitive to temperature: doubling an object's absolute temperature increases its radiated power per unit area sixteenfold. Real objects are not perfect blackbodies; a surface's actual radiated power is $P/A = \varepsilon \sigma T^4$, where the **emissivity** $\varepsilon \leq 1$ measures how closely the surface approximates an ideal absorber and emitter ($\varepsilon = 1$ for a true blackbody).

Worked Example: The Sun as a Blackbody

The Sun's surface (photosphere) has an effective temperature of $T \approx 5778$ K and radius $R_\odot = 6.96 \times 10^8$ m. Treating it as an ideal blackbody, Wien's law gives the wavelength of peak emission,

$$\lambda_{\max} = \frac{2.898 \times 10^{-3}\ \text{m}\cdot\text{K}}{5778\ \text{K}} = 5.02 \times 10^{-7}\ \text{m} = 502\ \text{nm}, \tag{6.7}$$

squarely in the green part of the visible spectrum — the Sun's spectrum actually peaks near the middle of the range of wavelengths the human eye evolved to detect, though scattering in Earth's atmosphere and the eye's overall spectral response make sunlight appear white or yellow rather than green. The Stefan–Boltzmann law gives the power radiated per unit area, $P/A = \sigma T^4 = (5.670 \times 10^{-8}\ \text{W}\cdot\text{m}^{-2}\cdot\text{K}^{-4})(5778\ \text{K})^4 \approx 6.32 \times 10^7\ \text{W/m}^2$. Multiplying by the Sun's surface area, $A = 4\pi R_\odot^2 \approx 6.09 \times 10^{18}\ \text{m}^2$, gives a total radiated power (luminosity)

$$P = \frac{P}{A} \cdot A \approx (6.32 \times 10^7\ \text{W/m}^2)(6.09 \times 10^{18}\ \text{m}^2) \approx 3.85 \times 10^{26}\ \text{W}, \tag{6.8}$$

in close agreement with the Sun's independently measured luminosity of 3.828×10^{26} W — the same figure used in Exercise 3.6 to estimate the Sun's rate of mass loss via $E = mc^2$. That two completely independent lines of physics (blackbody radiation here, and mass–energy equivalence there) connect so tightly to the same measured number is a testament to how well both theories describe the physical world.

Historical Context: Lenard's Puzzle

Planck's quantization hypothesis was aimed narrowly at the blackbody spectrum, and Planck himself did not initially believe that light itself came in discrete packets. Evidence pointing in that direction had, in fact, already been gathering independently. In 1902, three years before Einstein's photon paper, Philipp Lenard studied the photoelectric effect using an arc lamp and a retarding-voltage apparatus much like the one described below, and found — to general puzzlement at the time — that the stopping voltage (and hence the photoelectrons' maximum kinetic energy) did not increase with the lamp's intensity, contrary to every expectation from Maxwell's wave theory of light, in which a more intense wave carries proportionally more energy to whatever absorbs it. Lenard had no explanation; the result simply sat as an unresolved anomaly in the physics literature for three years, alongside blackbody radiation, until Einstein showed that both puzzles have the same root cause: light exchanges energy with matter in discrete quanta, not continuously.

6.2 The Photoelectric Effect and X-Ray Production

The Photoelectric Effect

When light of sufficiently short wavelength strikes a metal surface, electrons are ejected — the **photoelectric effect**. A typical apparatus measures the photocurrent as a function of a retarding voltage V applied between the emitting surface and a collector; the voltage at which the current just reaches zero, the **stopping potential** V_0, gives the maximum kinetic energy of the ejected electrons, $K_{\max} = eV_0$. Plotting the measured photocurrent against retarding voltage traces out a curve that falls smoothly to zero at V_0 rather than dropping abruptly,

because the ejected electrons are emitted with a distribution of kinetic energies up to $K_{\max}$, not a single sharp value; V_0 marks where even the *fastest* electrons are turned back.

Three experimental features of this effect resist any explanation in terms of classical electromagnetic waves, in which energy is delivered continuously and is proportional to intensity:

1. $K_{\max}$ **is independent of light intensity.** Classically, a more intense wave delivers more energy per unit time to an electron and should eject electrons with more kinetic energy. Experimentally, increasing intensity increases the *number* of photoelectrons (the current) but not $K_{\max}$.

2. $K_{\max}$ **depends linearly on frequency**, and there exists a sharp **threshold frequency** f_0, characteristic of the metal, below which no photoelectrons are emitted at all, regardless of intensity or exposure time. Classically, a wave of any frequency should eventually eject electrons if given enough time to deliver sufficient energy, so a hard threshold — and one depending on frequency rather than intensity — has no classical explanation.

3. **Emission is (essentially) instantaneous**, with no observable time lag even at very low intensity. Classically, a dim wave should take a measurable time to deliver enough energy to an electron to free it; a rough classical estimate for a very weak source predicts delays of minutes to hours, yet no such delay is ever observed.

Warning

A common misconception is to picture a brighter beam as made of more energetic photons, as if turning up a lamp's intensity somehow "speeds up" each photon. Intensity instead measures the *number* of photons arriving per second; each photon's energy hf is fixed by the light's frequency alone and is completely indifferent to how bright the source is. That is exactly why $K_{\max}$ tracks frequency rather than intensity: doubling the intensity doubles the photocurrent (twice as many photoelectrons ejected per second) but leaves the kinetic energy of each individual electron unchanged.

Einstein resolved all three in 1905 by proposing that light itself is quantized [Einstein, 1905a]: it consists of discrete packets, **photons**, each carrying energy

$$E = hf, \tag{6.9}$$

with h the same constant Planck introduced. In the photoelectric effect, a single photon transfers its entire energy to a single electron in one interaction. If ϕ (the metal's **work function**) is the minimum energy needed to remove an electron from the metal, conservation of energy gives

$$K_{\max} = hf - \phi, \tag{6.10}$$

the **photoelectric equation**. This immediately explains all three observations: $K_{\max}$ depends on f (through the photon energy hf) but not on intensity, since intensity only changes the *number* of photons per second, not the energy of each one; the threshold frequency is $f_0 = \phi/h$, below which a single photon simply does not carry enough energy to free an electron, no matter how many photons arrive; and emission is instantaneous because each electron absorbs one photon's energy all at once, not gradually. A plot of $K_{\max}$ (equivalently, eV_0) versus f is a straight line of slope h and y-intercept $-\phi$, and Millikan's precise measurement of

exactly this line (1916) provided both a direct experimental value of h and strong confirmation of Einstein's photon hypothesis — for which, not for relativity, Einstein received the 1921 Nobel Prize in Physics.

Note

Millikan spent nearly a decade trying to disprove Einstein's photon hypothesis, not confirm it. As a committed believer in the classical wave theory of light, he found the idea of discrete light quanta implausible and set out to show the photoelectric equation wrong experimentally. His painstaking measurements instead reproduced $K_{\max} = hf - \phi$ to within about half a percent across several metals, giving the first precise value of h from a source entirely independent of Planck's blackbody-spectrum fit. Millikan later described the result as "the enormously improbable equation which I was unable to evade," and won his own Nobel Prize in 1923, in part for exactly this measurement.

The three stubborn facts are best met in the order Lenard and Millikan met them, with the apparatus in front of you. In Figure 6.2 the light's intensity, its wavelength, the target metal, and the retarding voltage are all under control, and the photocurrent and stopping potential are read off directly. Turn the intensity up at fixed wavelength: the current rises and the stopping potential does not move. Shorten the wavelength instead: the stopping potential climbs, and a plot of eV_0 against f — the simulation will accumulate one for you — is a straight line whose slope is h and whose intercept names the metal. Cross the threshold from the long-wavelength side and the current stops altogether, at full intensity.

Tip

Before computing $K_{\max} = hf - \phi$, check whether the photon energy actually exceeds the work function — a negative result signals light below the threshold frequency (no photoelectrons at all), not a negative kinetic energy. Working consistently in electron-volts, with the shortcut $hc = 1240\ \text{eV} \cdot \text{nm}$, avoids most of the unit-conversion slips these problems invite; convert to joules or SI units only at the very end, if at all.

Worked Example: The Photoelectric Effect in Cesium

Cesium has one of the lowest work functions of any metal, $\phi = 2.14$ eV, which is why it is used in the photocathodes of photomultiplier tubes and early photoelectric light meters.

One electron-volt, $1\ \text{eV} = 1.602 \times 10^{-19}$ J, is the kinetic energy gained by a single electron accelerated through a 1 V potential difference — far more convenient than the joule for atomic-scale quantities like photon energies and work functions, which is why this chapter reports them in eV throughout.

Ultraviolet light of wavelength $\lambda = 250$ nm illuminates a cesium surface. Using the convenient combination $hc = 1240\ \text{eV} \cdot \text{nm}$,

$$E_{\text{photon}} = \frac{hc}{\lambda} = \frac{1240\ \text{eV} \cdot \text{nm}}{250\ \text{nm}} = 4.96\ \text{eV}. \tag{6.11}$$

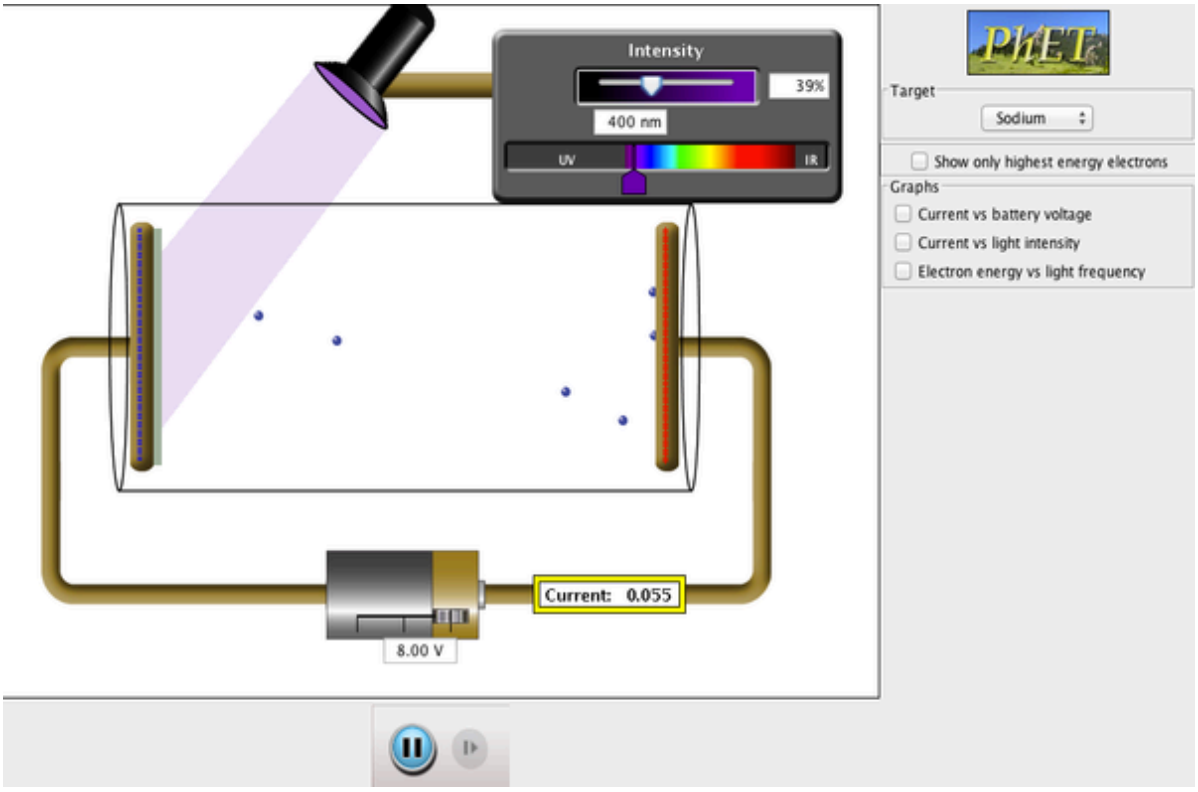

Figure 6.2: The photoelectric apparatus: a photocathode, an adjustable light source, and a retarding voltage. (This is one of PhET's original Java simulations, run in the browser by CheerpJ; it downloads a Java runtime before it starts, so give it a few seconds on first load.) Interactive simulation: Photoelectric Effect

The maximum kinetic energy of the photoelectrons is $K_{\max} = E_{\text{photon}} - \phi = 4.96\ \text{eV} - 2.14\ \text{eV} = 2.82\ \text{eV}$, so the stopping potential is $V_0 = 2.82\ \text{V}$. The threshold wavelength — the longest wavelength that can still eject an electron — follows from $\phi = hc/\lambda_0$:

$$\lambda_0 = \frac{hc}{\phi} = \frac{1240\ \text{eV} \cdot \text{nm}}{2.14\ \text{eV}} = 579\ \text{nm}, \tag{6.12}$$

which falls in the visible (yellow) part of the spectrum. Cesium is therefore photoelectrically sensitive to ordinary visible light, unlike most metals, whose work functions of 4–5 eV push their threshold wavelengths into the ultraviolet.

Applications: Photomultipliers and Photovoltaic Cells

The photoelectric effect underlies two technologies central to modern experimental physics and everyday life. A **photomultiplier tube** exploits a low-work-function photocathode (often a cesium compound, as in the worked example above) to convert a single incoming photon into a single ejected photoelectron, then accelerates that electron into a series of intermediate electrodes (**dynodes**), each of which is struck hard enough to eject several additional electrons by **secondary emission**; the resulting cascade, multiplying the electron count by a factor of a few at each of perhaps ten dynode stages, converts the arrival of a single photon into a macroscopic, easily measured current pulse containing millions of electrons. This single-photon sensitivity makes photomultiplier tubes essential wherever extremely faint light must be detected and counted, including the gamma-ray detectors used in PET scanners (discussed later in this chapter) and in nuclear and particle-physics experiments generally.

A closely related but distinct effect, the **photovoltaic effect**, underlies solar cells: rather than ejecting an electron entirely from the material, an absorbed photon in a semiconductor promotes an electron across the material's band gap (the semiconductor analog of the work function, examined further in solid-state contexts), and a built-in electric field at a junction between two differently doped semiconductor layers then separates the resulting electron and the vacancy (hole) it leaves behind before they can recombine, driving a current through an external circuit. As in the ordinary photoelectric effect, only photons with energy exceeding a threshold (the band-gap energy) contribute usable electrons, which is why solar-cell efficiency depends sensitively on matching the semiconductor's band gap to the solar spectrum.

X-Ray Production: Bremsstrahlung

The photoelectric effect converts a photon into an energetic electron; an energetic electron can conversely give up part of its energy as radiation when it is abruptly decelerated in matter. This **bremsstrahlung** process is the standard laboratory and industrial method of producing X-rays. It is an energy-bookkeeping inverse of the photoelectric effect, not its literal microscopic time reverse (which would involve radiative recombination into a bound state). In an X-ray tube, electrons are accelerated from rest through a large potential difference V (typically tens to hundreds of kilovolts) and strike a dense metal target (commonly tungsten). As each electron decelerates within the target — bremsstrahlung, German for "braking radiation" — it radiates part or all of its kinetic energy as one or more photons.

Because an individual electron can, in principle, lose its *entire* kinetic energy eV in a single deceleration event (converting essentially all its kinetic energy into one photon), there is a sharply defined **maximum** photon energy — and correspondingly a **minimum** wavelength — that the tube can produce, set by

$$eV = \frac{hc}{\lambda_{\min}} \quad \Longrightarrow \quad \lambda_{\min} = \frac{hc}{eV}, \tag{6.13}$$

known as the **Duane–Hunt limit**. Most electrons instead lose their energy gradually over many collisions, radiating photons of smaller energy at each step and producing a continuous spectrum of wavelengths longer than $\lambda_{\min}$ — but no photon in the entire spectrum can have a wavelength shorter than $\lambda_{\min}$, since no single electron carries more than eV of kinetic energy to begin with. Superimposed on this continuous bremsstrahlung spectrum are sharp **characteristic X-ray lines**, produced when an incident electron knocks an inner-shell electron out of a target atom and an outer electron falls down to fill the vacancy, emitting a photon of energy fixed by the target element's inner-shell energy levels (Moseley's law for these characteristic energies is developed in Chapter 11). Unlike the continuous bremsstrahlung background, the characteristic lines depend on the target material, not on the accelerating voltage (once V is large enough to eject the inner-shell electron in the first place).

Bremsstrahlung is worth separating into its classical and its quantum halves, because only the second is new here. The classical half — that a charge which accelerates radiates, and that the radiation carries energy away — is already in Maxwell's equations, and Figure 6.3 is that statement on its own: shake a charge and kinks in its field propagate outward at c, taking energy with them. Stopping a charge dead is an acceleration like any other, so an electron slamming into a tungsten anode must radiate. What classical physics cannot supply is the sharp edge at $\lambda_{\min}$: a continuous field theory sets no floor on the wavelength radiated in a single event, and the Duane–Hunt limit exists only because the radiated energy comes in quanta hf and one electron brings only eV to spend.

Worked Example: Minimum Wavelength from a Diagnostic X-Ray Tube

A medical diagnostic X-ray tube is operated at an accelerating voltage of $V = 80.0$ kV. The Duane–Hunt limit gives

$$\lambda_{\min} = \frac{hc}{eV} = \frac{1240\ \text{eV}\cdot\text{nm}}{8.00\times10^{4}\ \text{eV}} = 1.55\times10^{-2}\ \text{nm} = 15.5\ \text{pm}. \tag{6.14}$$

This is comparable to, and somewhat shorter than, typical interatomic spacings in crystals ($\sim$ 0.1–0.3 nm), which is why X-rays of this energy scale are useful for crystallographic diffraction (Chapter 5) as well as medical imaging, where their short wavelength (and correspondingly high photon energy) allows them to penetrate soft tissue while being partially absorbed by denser bone.

6.3 Compton Scattering, Pair Production, and the Photon

Compton Scattering

Even after the photoelectric effect, one could imagine "photon-like" energy exchange as a property specific to bound electrons in a metal, without light itself consisting of localized particles carrying momentum. Arthur Compton's 1923 experiment removed this loophole by showing that photons scattering from a *free* electron transfer momentum exactly as a particle collision would.

In Compton's experiment, X-rays of a single wavelength λ are directed at a target of loosely bound (effectively free) electrons, and the wavelength λ' of the scattered X-rays is measured

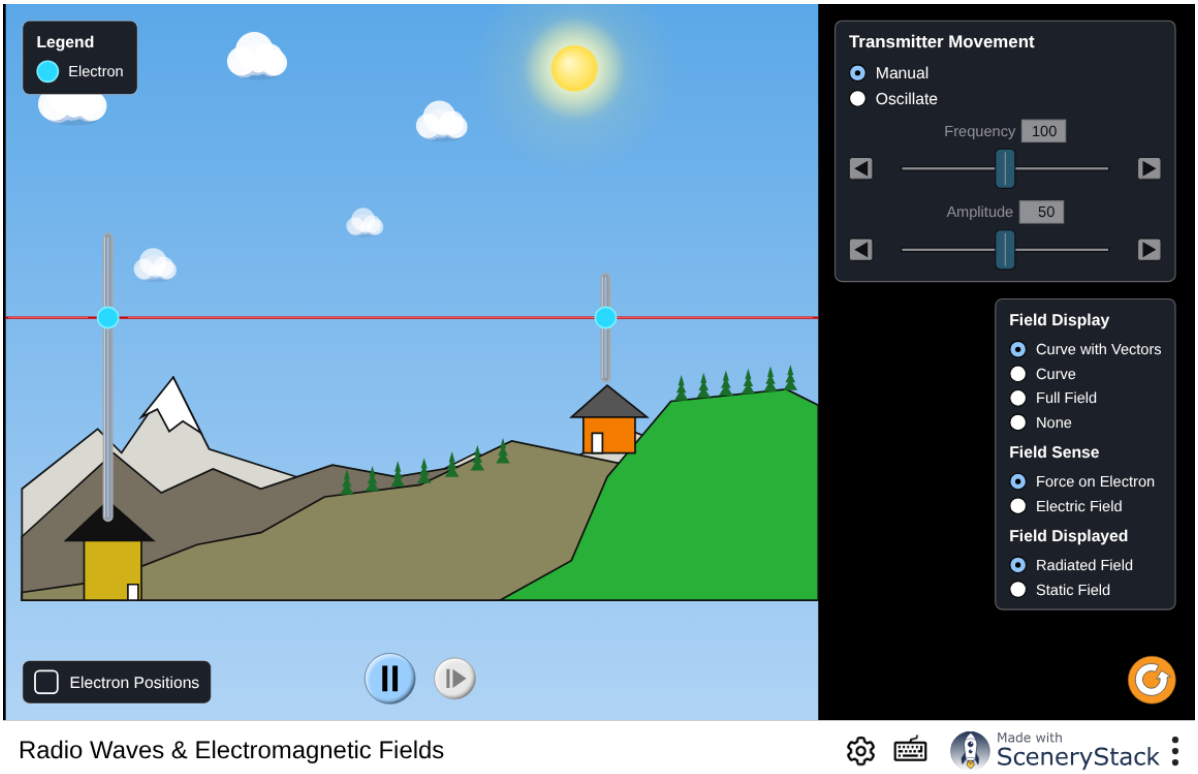

Figure 6.3: An accelerating charge and the field it radiates. Move the charge by hand and watch the disturbance propagate outward at c — the classical mechanism behind bremsstrahlung, and the one the electron in an X-ray tube obeys on its way to producing a photon of energy up to eV. Interactive simulation: Radio Waves

as a function of scattering angle θ. Classically, an electromagnetic wave incident on a charge should simply drive that charge to oscillate at the incident frequency and re-radiate at the *same* frequency (Thomson scattering); no wavelength shift is expected. Compton observed a systematic *increase* in wavelength, $\lambda' > \lambda$, growing with scattering angle θ and independent of the target material — a signature of a two-body collision, not wave re-radiation.

Deriving the Compton Formula

Treat the photon as a particle with energy $E = hc/\lambda$ and momentum $p = E/c = h/\lambda$ (consistent with the massless-particle limit of the energy–momentum relation from Chapter 3), and apply conservation of relativistic energy and momentum to an elastic collision between the photon and an initially free, stationary electron of mass m_e. Let the photon scatter through angle θ, emerging with wavelength λ', while the electron recoils with momentum p_e and (relativistic) energy E_e. The three momentum components entering the collision — incident photon momentum h/λ along the initial direction, and zero for the electron — must balance the two outgoing momenta, whose vector sum (photon momentum h/λ' at angle θ, electron momentum p_e at some recoil angle) forms a triangle. The law of cosines applied to that triangle gives

$$(p_e c)^2 = \left(\frac{hc}{\lambda}\right)^2 + \left(\frac{hc}{\lambda'}\right)^2 - 2\left(\frac{hc}{\lambda}\right)\left(\frac{hc}{\lambda'}\right)\cos\theta. \tag{i}$$

Conservation of energy, with the electron initially at rest ($E_e^{(0)} = m_e c^2$), gives

$$E_e = \frac{hc}{\lambda} - \frac{hc}{\lambda'} + m_e c^2. \tag{ii}$$

Squaring (ii) and using the energy–momentum invariant $E_e^2 = (p_e c)^2 + (m_e c^2)^2$ from Chapter 3 to eliminate E_e in favor of $p_e c$, then substituting (i) for $(p_e c)^2$, produces (after the $\left(\frac{hc}{\lambda}\right)^2$, $\left(\frac{hc}{\lambda'}\right)^2$, and $(m_e c^2)^2$ terms cancel identically between the two sides) the much simpler relation

$$m_e c^2\left(\frac{hc}{\lambda} - \frac{hc}{\lambda'}\right) = \frac{hc}{\lambda}\cdot\frac{hc}{\lambda'}(1 - \cos\theta). \tag{6.15}$$

Dividing through by hc, multiplying both sides by $\lambda\lambda'/m_e c^2$, and simplifying $\lambda\left(\frac{1}{\lambda} - \frac{1}{\lambda'}\right)\lambda' = \lambda' - \lambda$ yields the **Compton scattering formula**:

$$\lambda' - \lambda = \frac{h}{m_e c}(1 - \cos\theta). \tag{6.16}$$

The constant $h/m_e c = 2.426 \times 10^{-12}$ m is the **Compton wavelength** of the electron. The formula correctly predicts zero shift at $\theta = 0$ (forward, undeflected "scattering") and maximum shift $2h/m_e c$ at $\theta = 180°$ (photon backscattered), matches the observed angular dependence precisely, and — crucially — is independent of the incident wavelength λ itself, matching experiment. Compton scattering is direct, quantitative confirmation that a photon carries momentum $p = h/\lambda$ and transfers it to a free electron exactly as one particle colliding with another.

The Compton formula in terms of photon energy

The wavelength form $\lambda' - \lambda = (h/m_e c)(1 - \cos\theta)$ is the natural one for X-ray diffraction work, where wavelengths are measured directly, but gamma-ray spectroscopy — including the Compton-edge discussion below — more often calls for the scattered photon's *energy* $E' =$

hc/λ' directly in terms of the incident energy $E = hc/\lambda$. Substituting $\lambda = hc/E$ and $\lambda' = hc/E'$ into the Compton formula gives

$$\frac{hc}{E'} - \frac{hc}{E} = \frac{h}{m_e c}(1 - \cos\theta). \tag{6.17}$$

Multiplying through by $EE'/(hc)$ and solving for E' yields

$$E' = \frac{E}{1 + \dfrac{E}{m_e c^2}(1 - \cos\theta)}. \tag{6.18}$$

Setting $\theta = 180°$ gives the *minimum* possible scattered-photon energy, $E'_{\min} = E/[1 + 2E/(m_e c^2)]$, and hence the *maximum* possible energy $\Delta E_{\max} = E - E'_{\min}$ that a single Compton-scattering event can deposit in a detector — the location of the Compton edge introduced below.

Worked Example: Compton-Scattered Molybdenum X-Rays

X-rays of wavelength $\lambda = 0.100$ nm (comparable to characteristic molybdenum K_α X-rays used in crystallography) Compton-scatter off free electrons at $\theta = 60°$. The wavelength shift is

$$\Delta\lambda = \frac{h}{m_e c}(1 - \cos 60°) = (2.426 \times 10^{-12}\ \text{m})(1 - 0.500) = 1.21 \times 10^{-12}\ \text{m} = 1.21\ \text{pm}, \tag{6.19}$$

so the scattered wavelength is $\lambda' = 101.2$ pm. In photon-energy terms, the incident photon carries $E = hc/\lambda = (1240\ \text{eV}\cdot\text{nm})/(0.100\ \text{nm}) = 12.40$ keV, while the scattered photon carries $E' = hc/\lambda' = (1240\ \text{eV}\cdot\text{nm})/(0.1012\ \text{nm}) = 12.25$ keV. The energy transferred to the recoiling electron is therefore $\Delta E = E - E' \approx 0.15$ keV $= 150$ eV — a small but entirely measurable fraction of the incident photon's energy, exactly the kind of energy loss Compton measured to confirm the formula.

The Compton Edge

In a real gamma-ray or X-ray detector, photons scatter through the full range of angles $0 \le \theta \le 180°$ available inside the detector material, and the detector records the energy $\Delta E(\theta) = E - E'(\theta)$ actually deposited by the recoiling electron for each scattering event. Because $\Delta E(\theta)$ increases monotonically with θ (least energy transfer for forward scattering, most for backscattering), there is a sharply defined *maximum* possible energy deposit, occurring at $\theta = 180°$ — the **Compton edge** — beyond which no Compton-scattered electron can deposit more energy in a single scattering event, no matter how many photons are examined. A photon that instead deposits its *entire* energy in one interaction (via the photoelectric effect on a bound atomic electron, rather than Compton scattering a free one) produces a separate, sharp peak at the full incident photon energy, the **photopeak**. Gamma-ray spectroscopists routinely distinguish these two features — a sharp photopeak plus a broad continuum of Compton-scattered energies cut off abruptly at the Compton edge — when interpreting a detector's measured energy spectrum, since both are simultaneous, competing ways the same photon can interact with the detector material.

Pair Production and Annihilation

A further, still more dramatic demonstration of the particle nature of light is **pair production**: a sufficiently energetic photon, passing near a nucleus, can convert entirely into an electron–positron pair,

$$\gamma \to e^- + e^+, \tag{6.20}$$

requiring a photon energy of at least $2m_e c^2 = 1.022$ MeV (twice the electron rest energy) — a direct manifestation of mass–energy equivalence (Chapter 3), converting a massless particle's energy into the rest mass of two massive particles.

A nearby massive third body — typically an atomic nucleus — is required to conserve momentum. This can be seen directly from the energy–momentum four-vector formalism of Chapter 3: an isolated photon of energy E_γ has momentum $p_\gamma = E_\gamma/c$, but a resulting electron–positron pair with the *same* total energy E_γ (by energy conservation) necessarily has total momentum strictly *less* than E_γ/c, because each massive particle satisfies $E^2 = (pc)^2 + (mc^2)^2 > (pc)^2$, so $pc < E$ for each — the pair's combined momentum cannot match the photon's original momentum at the same total energy. A nucleus nearby can absorb the small difference in momentum (recoiling with negligible kinetic energy, because its mass is so much larger than m_e) while barely affecting the energy balance, resolving the mismatch; in a vacuum with no such third body available, momentum conservation alone forbids the process outright, regardless of how much energy the lone photon carries.

The reverse process, **pair annihilation**, $e^- + e^+ \to 2\gamma$ (two photons are required, rather than one, to conserve momentum in the electron–positron center-of-momentum frame — see Chapter 3, Exercise 3.3), converts rest mass entirely back into photon energy and is used, for example, in positron-emission tomography (PET) imaging, where each annihilation of an injected positron-emitting tracer with a nearby atomic electron produces two back-to-back 511 keV gamma rays that a ring of detectors uses to reconstruct the tracer's location.

> **See Also**
> The radioactive decays that produce gamma rays in the first place, and the nuclear physics behind tracers like the ones PET imaging relies on, are developed in Chapter 13, which picks up where this chapter's photon–matter interactions leave off.

Worked Example: Sharing Energy Above Pair-Production Threshold

A photon of energy $E_\gamma = 3.00$ MeV, well above the 1.022 MeV threshold, undergoes pair production near a heavy nucleus. Because the nucleus is far more massive than the electron or positron, it absorbs essentially none of the available kinetic energy (a large mass can supply whatever small momentum balance is needed while carrying away almost no energy, exactly analogous to how a wall barely recoils, and gains almost no kinetic energy, when it reflects a ball). The energy left over after paying the $2m_e c^2$ rest-mass cost is therefore shared, to good approximation, between the electron and positron as kinetic energy:

$$K_{e^-} + K_{e^+} \approx E_\gamma - 2m_e c^2 = 3.00 \text{ MeV} - 1.022 \text{ MeV} = 1.978 \text{ MeV}. \tag{6.21}$$

If the pair shares this energy symmetrically (not required by any conservation law, but a useful simplifying assumption), each particle carries about 0.989 MeV of kinetic energy —

comparable to its own rest energy, so each is produced at a substantial fraction of the speed of light.

The Photon: A Unified Picture

Four seemingly unrelated phenomena — blackbody radiation, the photoelectric effect, X-ray production, and Compton scattering — together with pair production and annihilation, are all consequences of a single underlying fact: electromagnetic radiation exchanges energy and momentum with matter in discrete, particle-like quanta, $E = hf$ and $p = h/\lambda$, rather than continuously. Blackbody radiation shows that an oscillator can only *emit or absorb* energy in these units; the photoelectric effect shows that a photon transfers its *entire* energy to a single electron in one step; X-ray production shows how an electron's kinetic energy can be converted into photons, with a sharply bounded maximum photon energy; Compton scattering shows that a photon carries not just energy but momentum, exchanged with a free electron exactly as in a two-body collision; and pair production and annihilation show that a photon's energy is, via $E = mc^2$, interconvertible with the rest mass of matter itself. The characteristic quantum features of all five are explained, quantitatively and without exception, by the same photon concept. Yet, as Chapters 4–5 demonstrated, light also produces interference and diffraction patterns with no possible explanation in a naive particle picture. Chapter 7 confronts this apparent contradiction directly, and shows that it is resolved not by choosing one description over the other, but by recognizing that matter, too, has a wave nature — and that wave and particle descriptions are two complementary faces of a single, more complete quantum picture.

6.4 Summary

- The blackbody spectrum cannot be explained by the classical **Rayleigh–Jeans law**, $u(\lambda, T) = 8\pi k_B T/\lambda^4$, which diverges at short wavelength (the ultraviolet catastrophe); Planck's hypothesis that oscillator energy is quantized in units of hf yields a radiation law matching observation at all wavelengths and reduces to the Rayleigh–Jeans law at long wavelength.
- **Wien's displacement law**, $\lambda_{\max} T = 2.898 \times 10^{-3}\ \mathrm{m \cdot K}$, and the **Stefan–Boltzmann law**, $P/A = \sigma T^4$, follow from Planck's radiation law and together determine both the color and total radiated power of a thermal source from its temperature alone.
- The photoelectric effect — $K_{\max}$ independent of intensity, a sharp threshold frequency, and instantaneous emission — cannot be explained by classical wave theory but follows directly from Einstein's photon hypothesis, $E = hf$, via $K_{\max} = hf - \phi$.
- X-ray production (bremsstrahlung) is an energy-bookkeeping inverse of the photoelectric effect: decelerating electrons in an X-ray tube produce a continuous spectrum bounded by a sharp minimum wavelength, the **Duane–Hunt limit**, $\lambda_{\min} = hc/eV$, since no photon can carry more energy than a single electron's kinetic energy eV.
- Compton scattering — the angle-dependent wavelength shift of X-rays scattered from free electrons, $\lambda' - \lambda = (h/m_e c)(1 - \cos\theta)$, derivable directly from conservation of relativistic energy and momentum — shows that photons carry momentum $p = h/\lambda$ and interact with electrons as particles in an elastic collision.
- Pair production ($\gamma \to e^- + e^+$, requiring $hf \geq 2m_e c^2$ and a nearby massive body to conserve momentum) and pair annihilation ($e^- + e^+ \to 2\gamma$) further confirm that photons are quanta of energy and momentum interconvertible with massive particles.

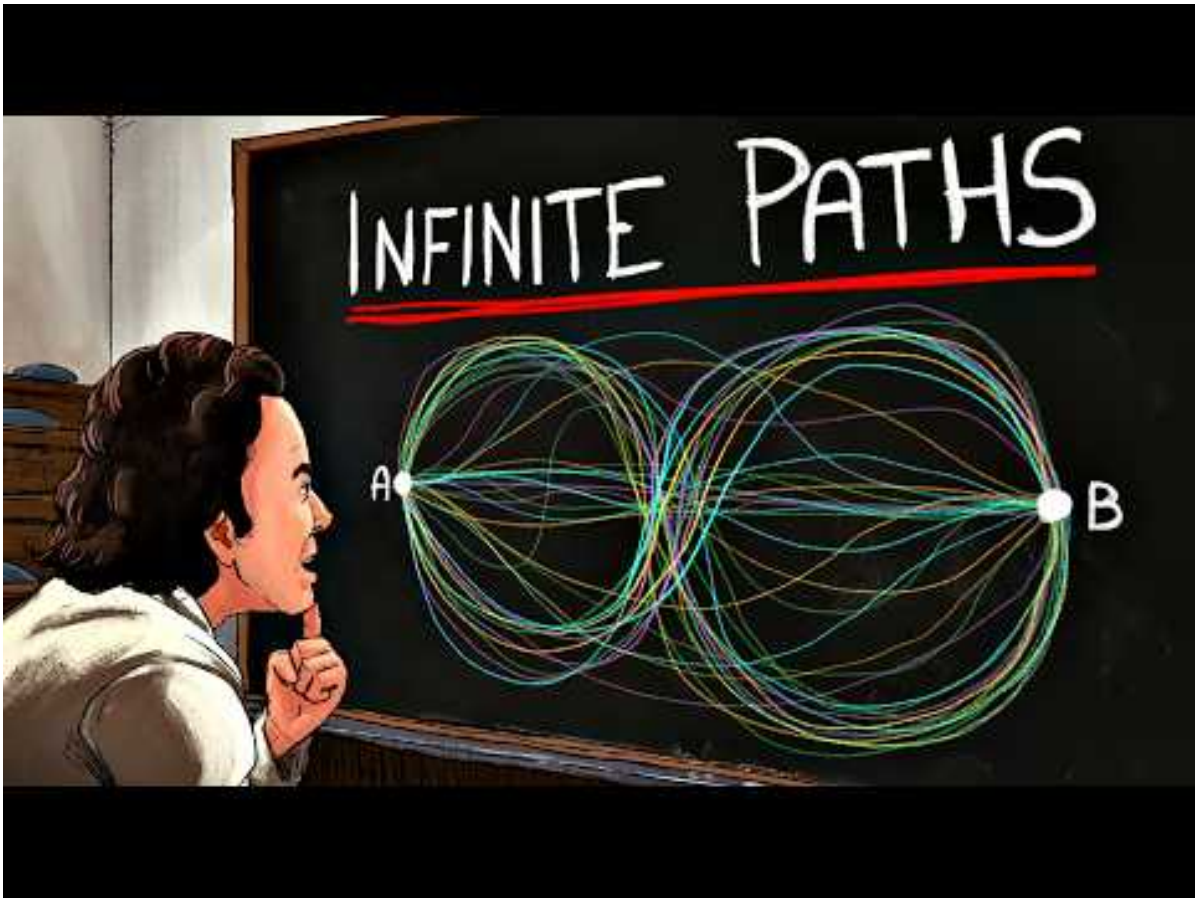

Figure 6.4: A closing look at just how far Einstein's photon hypothesis — introduced in this chapter for the photoelectric effect — ends up reaching. Watch on YouTube: The Strangest Implication of Quantum Mechanics

6.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats.

1. **Multiple choice.** Which observation most directly shows that photoelectric energy arrives in quanta? Choices: (A) Below a threshold frequency no electrons are emitted, regardless of intensity. (B) Brighter light always gives each emitted electron more kinetic energy. (C) All wavelengths eject electrons if the exposure lasts long enough. (D) The stopping potential depends only on the illuminated area.

2. **True or false.** At fixed frequency above threshold, increasing light intensity raises the maximum kinetic energy of photoelectrons.

3. **Drag the words.** Complete the photon relations. A photon has energy _____ and momentum _____. Photoemission obeys _____. In Compton scattering, a photon that loses energy has a _____. (Terms: E = hf; p = h/λ; Kmax = hf - ϕ; longer wavelength; E = h/f; shorter wavelength.)

4. **Fill in the blanks.** Complete the pair-production and thermal-radiation results. The rest-energy threshold for electron–positron pair production is _____ MeV, and total blackbody power per area scales as temperature to the power _____.

5. **Mark the words.** Mark the two observations that support the photon model. “The photoelectric effect has a sharp threshold frequency, and Compton scattering shows an angle-dependent wavelength shift. Classical intensity alone does not explain either result.”

6.6 Conceptual Questions

1. Why did the ultraviolet catastrophe show that classical equipartition could not describe blackbody radiation?

2. Increasing a light beam’s intensity increases photoelectric current but not the stopping potential. Explain both observations with photons.

3. Why does the Compton shift depend on scattering angle but not on the intensity of the incident beam?

4. X-ray tubes convert only part of an electron’s kinetic energy into one photon. Why is there nevertheless a sharp minimum wavelength?

5. Why can a single photon not produce an electron–positron pair in empty space while conserving both energy and momentum?

6. In what sense do the photoelectric and Compton effects reveal different particle-like properties of light?

6.7 Problems

Exercise 6.1
(Straightforward)

The peak of the Sun's blackbody emission spectrum occurs near $\lambda_{\max} \approx 500$ nm. Use Wien's displacement law to estimate the Sun's surface temperature, and compare your answer to the value $T = 5778$ K used in the worked example.

Exercise 6.2
(Straightforward)

A tungsten filament in an incandescent bulb operates at $T = 2900$ K. (a) Find the wavelength of peak emission and identify the region of the spectrum (visible, infrared, etc.) in which it lies. (b) Explain, using your answer to (a), why incandescent bulbs are inefficient sources of visible light. (c) Find the power radiated per unit area, assuming ideal blackbody behavior.

Exercise 6.3
(Straightforward)

A red giant star has surface temperature 3200 K and radius $500\,R_\odot$, where $R_\odot = 6.96 \times 10^8$ m is the Sun's radius. Using the Stefan–Boltzmann law, find the star's total luminosity, and compare it (as a ratio) to the Sun's luminosity of 3.83×10^{26} W found in the worked example.

Exercise 6.4
(Straightforward)

Light of wavelength 400 nm strikes a sodium surface with work function $\phi = 2.28$ eV. (a) Find the photon energy in eV. (b) Find the maximum kinetic energy of the ejected photoelectrons. (c) Find the stopping potential V_0. (d) Find the threshold wavelength for sodium.

Exercise 6.5
(Straightforward)

A photoelectric-effect experiment on a certain metal gives a stopping potential of 0.65 V for light of wavelength 450 nm, and 1.28 V for light of wavelength 360 nm. Use these two data points (rather than assuming a value of h) to determine (a) Planck's constant and (b) the work function of the metal from this data, treating $K_{\max} = eV_0 = hf - \phi$ as a linear equation in f.

Exercise 6.6
(Straightforward)

A photocell with a platinum surface (work function $\phi = 6.35$ eV) is illuminated with light of wavelength 150 nm. (a) Show that this wavelength is above the threshold for photoemission and find $K_{\max}$. (b) Find the longest wavelength of light that could eject photoelectrons from platinum, and explain why ordinary visible or near-UV light sources cannot do so.

Exercise 6.7
(Straightforward)

An X-ray tube is operated at an accelerating voltage of 120 kV. (a) Find the Duane–Hunt minimum wavelength $\lambda_{\min}$. (b) Find the corresponding maximum photon energy, in keV. (c) If the accelerating voltage is doubled, by what factor does $\lambda_{\min}$ change?

Exercise 6.8

(Straightforward)

Explain, in your own words and using conservation of energy, why the bremsstrahlung spectrum from an X-ray tube is continuous (a range of wavelengths) even though each individual electron carries the same, sharply defined kinetic energy eV upon striking the target.

Exercise 6.9

(Moderate)

X-rays of wavelength $\lambda = 0.0711$ nm are Compton-scattered at $\theta = 90°$. (a) Find the wavelength shift. (b) Find the wavelength and energy of the scattered photon. (c) Find the kinetic energy given to the recoiling electron.

Exercise 6.10

(Moderate)

Show that, for a photon Compton-scattered directly backward ($\theta = 180°$) by a free electron, the wavelength shift is $2h/m_e c$, and evaluate this numerically. For an incident photon of very short wavelength ($\lambda \ll h/m_e c$, i.e. a very energetic photon), find the approximate fraction of the incident photon's energy that is transferred to the electron in a 180° collision.

Exercise 6.11

(Moderate)

Fill in the algebraic step omitted in the text: starting from equations (i) and (ii) in the derivation of the Compton formula, substitute (i) into the squared version of (ii) combined with $E_e^2 = (p_e c)^2 + (m_e c^2)^2$, and verify explicitly that the terms $(hc/\lambda)^2$, $(hc/\lambda')^2$, and $(m_e c^2)^2$ cancel between the two sides, leaving the simplified relation quoted in the text.

Exercise 6.12

(Moderate)

A beam of visible-light photons ($\lambda = 600$ nm) is Compton-scattered from free electrons at $\theta = 90°$. (a) Compute the fractional wavelength shift $\Delta\lambda/\lambda$. (b) Explain, using your numerical result, why the Compton effect is essentially undetectable with visible light and was only discovered using X-rays.

Exercise 6.13

(Moderate)

Find the minimum photon energy required for pair production of an electron-positron pair, and convert this to a wavelength. Explain why pair production cannot occur for an isolated photon in empty space (i.e., why a nearby nucleus or other particle is required), using conservation of momentum and energy together.

Exercise 6.14

(Moderate)

A 2.50 MeV photon undergoes pair production near a heavy nucleus, which recoils with negligible kinetic energy. If the electron and positron share the available kinetic energy equally, find the kinetic energy and the total (kinetic plus rest) energy of each particle.

Exercise 6.15
(Moderate)

In positron-emission tomography, a positron emitted by a radioactive tracer travels a short distance, loses its kinetic energy through collisions, and then annihilates with an atomic electron essentially at rest. (a) Using conservation of energy and momentum (as in Exercise 3.3), find the energy of each of the two emitted gamma rays. (b) Find the wavelength of each gamma ray. (c) Explain, physically, why PET scanners are built as a ring of detectors surrounding the patient rather than a single detector.

Exercise 6.16
(Moderate)

Estimate the minimum photon energy needed to produce a proton–antiproton pair ($p + \bar{p}$) instead of an electron–positron pair, given the proton rest energy $m_p c^2 = 938$ MeV, and compare the required photon wavelength to that found for electron–positron pair production in Problem 13.

Exercise 6.17
(Challenging)

A 0.662 MeV gamma ray (a characteristic energy emitted by the common radioactive source cesium-137) Compton-scatters inside a detector. (a) Find the maximum possible energy, $\Delta E_{\max}$, that can be deposited by a single Compton-scattering event (i.e., the location of the Compton edge), using the $\theta = 180°$ result from Problem 10. (b) Find the photopeak energy, i.e., the energy deposited if the photon instead undergoes full photoelectric absorption. (c) Explain, in one or two sentences, why a real detector's measured spectrum for this source shows both features.

Exercise 6.18
(Challenging)

Explain, using the discussion of photomultiplier tubes, why a single visible-light photon (energy of order a few eV) can trigger a measurable macroscopic current pulse, even though a single electron's charge (1.6×10^{-19} C) is far too small to detect directly with ordinary circuitry; estimate the number of electrons that must reach the final dynode stage to constitute a charge pulse of 1.0×10^{-12} C (a typical minimum detectable pulse), and comment on whether a cascade of ten dynode stages, each multiplying the electron count by a factor of 4, is sufficient to produce a pulse of this size starting from one photoelectron.

Exercise 6.19
(Challenging)

Explain why the ultraviolet catastrophe (Rayleigh–Jeans law) and Lenard's photoelectric puzzle are, at first glance, very different experimental problems (one about the color and intensity of thermal glow, the other about ejected electrons), yet both are resolved by exactly the same hypothesis. Identify precisely which physical assumption is shared by both classical (failed) treatments, and which single quantization postulate replaces it in both cases.

Exercise 6.20

(Challenging)

Silicon, used in most solar cells, has a band gap of about 1.1 eV. (a) Find the longest wavelength of light that can produce a usable electron–hole pair in a silicon photovoltaic cell, and identify the region of the spectrum in which it lies. (b) Explain, using the photoelectric equation as a guide, why photons with much *more* than the band-gap energy do not produce proportionally more usable electrical energy per photon, even though they carry more energy each.

Chapter 7

Wave Properties of Particles

Learning Objectives

By the end of this chapter, you should be able to:

- State the de Broglie hypothesis and compute the de Broglie wavelength of a particle.
- Explain how the de Broglie hypothesis reinterprets the Bohr model's angular-momentum quantization condition as a standing-wave requirement.
- Describe the Davisson–Germer and G. P. Thomson experiments and explain why they confirm the de Broglie hypothesis, and describe how matter-wave interference has since been extended to atoms and large molecules.
- Explain wave–particle duality and complementarity, and describe how "which-path" information destroys interference.
- Construct a wave packet as a superposition of waves and relate its spatial and momentum spread; distinguish phase velocity from group velocity and show that group velocity equals particle velocity.
- State, derive qualitatively, and apply the Heisenberg uncertainty principle for position–momentum and energy–time.
- Use the uncertainty principle to estimate minimum energies and to explain why electrons cannot exist inside a nucleus.

Introduction

Chapter 6 established that light, long understood as a wave, also behaves as a stream of particle-like quanta in its interactions with matter. In 1924, Louis de Broglie proposed the converse: that matter, long understood as composed of particles, should also exhibit wave behavior. This was not an experimental discovery but a bold symmetry argument, made in de Broglie's doctoral thesis, and it proved correct. This chapter develops the de Broglie hypothesis, the experiments that confirmed it, and its most important consequence: because a particle with a well-defined wavelength is necessarily spread out in space, position and momentum cannot both be known with unlimited precision. That trade-off, the Heisenberg uncertainty principle, is not a statement about the limits of measurement technique but a fundamental feature of nature, and it sets the stage for the wave mechanics developed in Chapter 8.

7.1 Matter Waves: de Broglie and Experiment

The de Broglie Hypothesis

Chapter 6 established that a photon carries momentum related to its wavelength by $p = h/\lambda$. De Broglie's proposal was to take this relation, turn it around, and apply it universally: **every material particle of momentum p has an associated wavelength**

$$\lambda = \frac{h}{p}, \tag{7.1}$$

now called the **de Broglie wavelength**. For a nonrelativistic particle of mass m and speed u, $p = mu$, so $\lambda = h/mu$; more generally, for a particle accelerated from rest through a potential difference V, $K = eV = p^2/2m$ gives $p = \sqrt{2meV}$ and hence $\lambda = h/\sqrt{2meV}$. Because h is so small, the de Broglie wavelength of ordinary macroscopic objects is utterly negligible — a 1 g mass moving at 1 m/s has $\lambda \sim 10^{-31}$ m, far too small to produce any observable wave effect — which is why matter waves went unnoticed until physicists deliberately looked for them in systems where λ is not negligible, such as low-energy electrons, whose small mass makes λ comparable to atomic and crystal-lattice spacings for accessible kinetic energies.

> $h = 6.626 \times 10^{-34}$ J · s, the same constant from Chapter 6's photon relations. It is often quoted instead as $\hbar \equiv h/2\pi = 1.055 \times 10^{-34}$ J · s, the form used later in this chapter for wave packets and the uncertainty principle.

Note

De Broglie's hypothesis first appeared not in a journal article but in his 1924 doctoral thesis, and his examiners at the Sorbonne were unsure what to make of a proposal with no direct experimental support. His advisor, Paul Langevin, sent a copy to Einstein for a second opinion; Einstein reportedly replied that de Broglie "has lifted a corner of the great veil" and urged that the thesis be taken seriously. De Broglie was awarded his doctorate, and the Nobel Prize in Physics followed just five years later, in 1929 — an unusually fast recognition for a hypothesis confirmed experimentally only in 1927 (see below).

Tip

Before computing a de Broglie wavelength, check whether the particle's kinetic energy K is small or large compared to its rest energy mc^2. If $K \ll mc^2$ (true for the 100 eV electrons below, since $m_e c^2 = 511$ keV), the nonrelativistic relations $p = mu$ or $p = \sqrt{2mK}$ are accurate; once K becomes comparable to or larger than mc^2 (as for MeV-scale electrons or GeV-scale protons), $\lambda = h/p$ still holds, but p must instead be found from the relativistic relation $E^2 = (pc)^2 + (mc^2)^2$.

Worked Example: De Broglie Wavelength of an Accelerated Electron

An electron is accelerated from rest through a potential difference of $V = 100$ V, a typical value in early electron-diffraction experiments. Its nonrelativistic momentum is

$$p = \sqrt{2m_e eV} = \sqrt{2(9.11 \times 10^{-31}\ \text{kg})(1.60 \times 10^{-19}\ \text{C})(100\ \text{V})} = 5.40 \times 10^{-24}\ \text{kg} \cdot \text{m/s}, \quad (7.2)$$

so the de Broglie wavelength is

$$\lambda = \frac{h}{p} = \frac{6.626 \times 10^{-34}\ \text{J} \cdot \text{s}}{5.40 \times 10^{-24}\ \text{kg} \cdot \text{m/s}} = 1.23 \times 10^{-10}\ \text{m} = 0.123\ \text{nm}. \quad (7.3)$$

This convenient combination is often written $\lambda[\text{nm}] \approx 1.226/\sqrt{V[\text{volts}]}$ for nonrelativistic electrons. The result, 0.123 nm, is comparable to the spacing between atomic planes in a crystal ($\sim$ 0.1–0.3 nm) — exactly the condition needed for the crystal to act as a diffraction grating for the electrons, as the next section describes.

A Consistency Check: The Relativistic Origin of $\lambda = h/p$

De Broglie did not simply guess $\lambda = h/p$; he arrived at it by demanding consistency between the photon relations of Chapter 6 and the relativistic energy–momentum relation of Chapter 3. A wave of frequency f and phase speed v_p has wavelength $\lambda = v_p/f$. Associating a particle of energy E with a wave of frequency $f = E/h$ (the Planck–Einstein relation, extended by hypothesis to matter) and demanding that the resulting phase velocity be consistent with the particle's momentum p via $E = pc^2/u$ (the relativistic relation between energy, momentum, and velocity established in Chapter 3) leads, after eliminating f and v_p in favor of E, p, and the particle's actual velocity u, directly to $\lambda = h/p$ — the same relation obtained by simply carrying the photon formula over to matter, but now derived from relativistic energy–momentum consistency rather than asserted by analogy alone. This is one reason de Broglie's hypothesis, though speculative, was taken seriously immediately: it was not an arbitrary guess but the unique relation consistent with treating matter and light on the same relativistic footing.

The Davisson–Germer and G. P. Thomson Experiments

Direct confirmation came in 1927, when Clinton Davisson and Lester Germer, studying electron scattering from a nickel crystal (originally for an unrelated purpose), observed that the intensity of electrons scattered from the crystal surface showed sharp maxima at specific angles, depending on the electrons' kinetic energy — exactly the pattern expected from **diffraction** of a wave by the regularly spaced planes of atoms in the crystal, analogous to X-ray diffraction from a crystal lattice (Bragg diffraction). Measuring the angles of the diffraction maxima and applying the same diffraction condition used for X-rays,

$$n\lambda = 2d \sin\theta, \quad (7.4)$$

(with d the crystal's known interplanar spacing and θ measured from the planes) allowed Davisson and Germer to extract an experimental wavelength for the electrons — and it agreed, to good precision, with the de Broglie wavelength $\lambda = h/p$ computed from the electrons' known kinetic energy [Davisson and Germer, 1927].

> This is the same **Bragg condition** used for X-ray diffraction from crystals: successive lattice planes reinforce a scattered wave only when the total path difference $2d\sin\theta$ between reflections off adjacent planes is a whole number of wavelengths. Davisson and Germer's

result is exactly this condition, satisfied by an electron's matter wave instead of an X-ray photon.

The experiment is reconstructed in Figure 7.1, where both sides of $n\lambda = 2d\sin\theta$ are under control: the electron gun's accelerating voltage, which fixes p and hence $\lambda = h/p$, and the spacing of the target atoms, which fixes d. Raising the voltage shortens the wavelength and pulls the scattering maxima to smaller angles; widening the atomic spacing pushes them back out. Neither response is available to a stream of particles bouncing off a surface, and observing both is what settled the question.

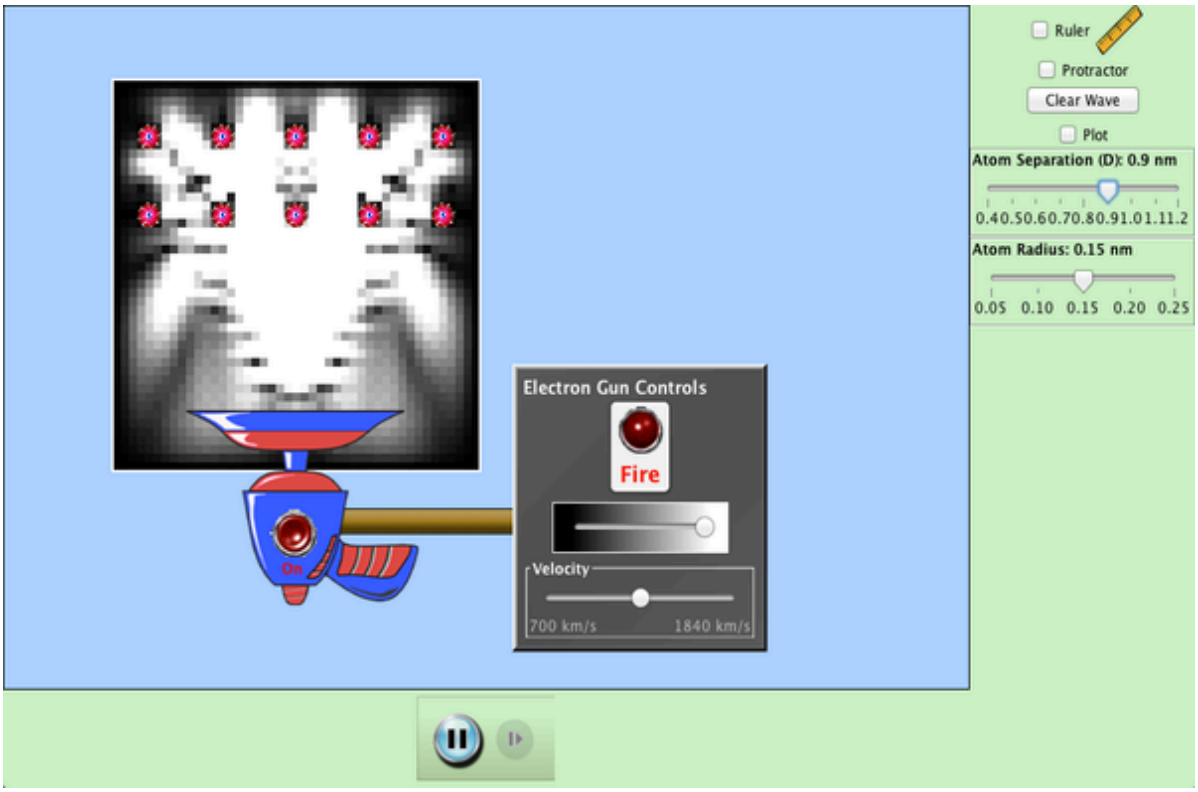

Figure 7.1: Electrons scattering from a crystal surface, with the beam energy and the atomic spacing under control. What is drawn is the electron's wave; where it is detected is a single point. Interactive simulation: Davisson–Germer: Electron Diffraction

That same year, working independently in Britain, George Paget Thomson fired higher-energy electrons through thin polycrystalline metal foils and observed concentric diffraction rings on a photographic plate — the electron analog of the ring patterns produced by X-ray diffraction through a powdered crystalline sample — providing an independent confirmation using a completely different experimental geometry. Thomson's result carries a particular irony: his father, J. J. Thomson, had won the 1906 Nobel Prize for discovering the electron and demonstrating that it is a particle with a definite charge-to-mass ratio; the son shared the 1937 Nobel Prize (with Davisson) for demonstrating that the very same particle also behaves as a wave — both experiments correct, neither in conflict with the other, once wave–particle duality is properly understood.

Electrons, unambiguously particles in every other respect (they have definite charge and mass, and leave localized tracks and point-like impacts on a detector), diffract like waves when

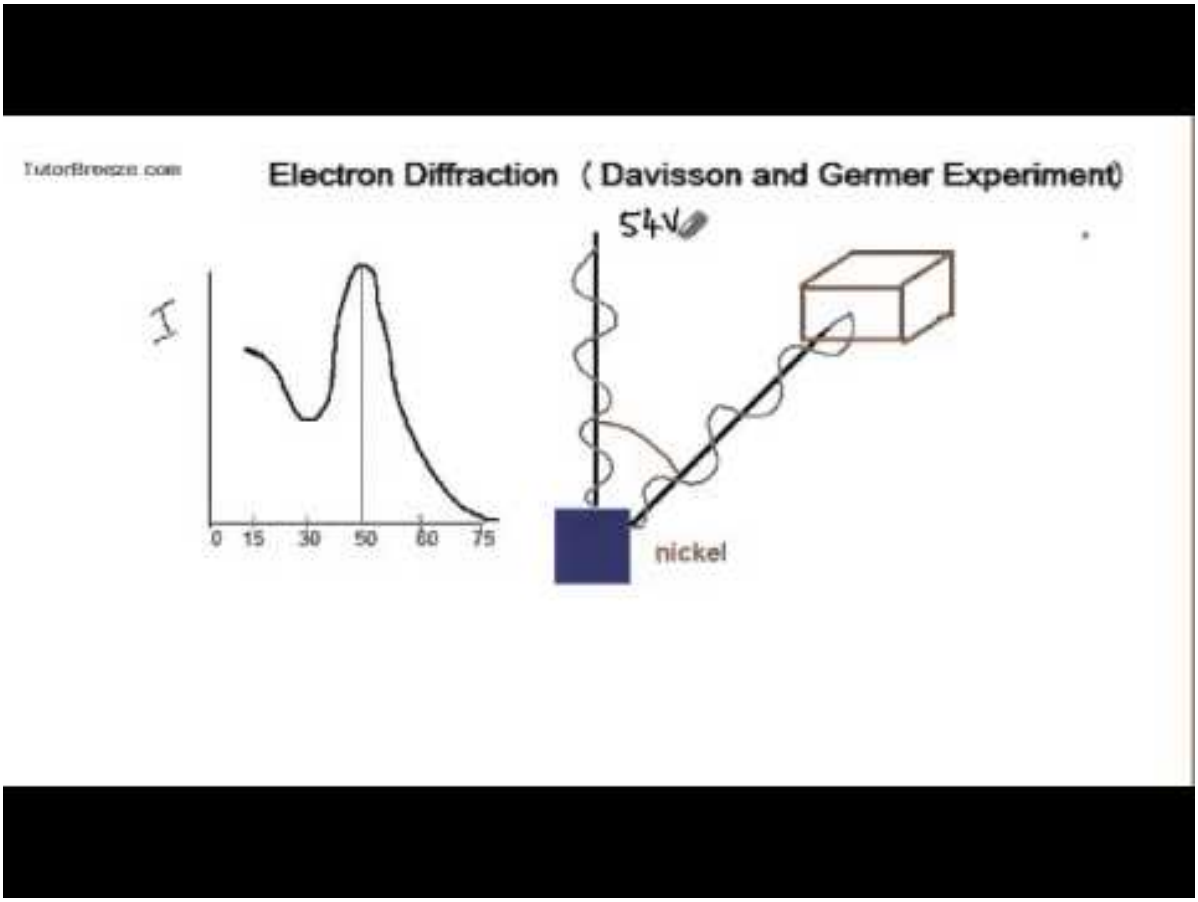

Figure 7.2: The historical apparatus and its intensity-versus-angle result, presented as the experimental bridge from de Broglie's hypothesis to measured electron diffraction. Watch on YouTube: Davisson and Germer Experiment — Electron Diffraction

their de Broglie wavelength is comparable to the spacing of the diffracting structure. The effect has since been confirmed for neutrons, atoms, and — in experiments beginning in the late 1990s — even large molecules: neutron diffraction is now a standard tool for probing crystal and magnetic structure (complementing X-ray diffraction, since neutrons, being uncharged, scatter primarily from nuclei rather than electron clouds), and a landmark 1999 experiment by Arndt, Zeilinger, and collaborators observed diffraction of C_{60} "buckyball" molecules — each containing 60 carbon atoms, with a mass roughly 10^6 times that of a single electron — through a microfabricated grating, with a measured de Broglie wavelength of only a few picometers, far smaller than the molecule itself, yet still large enough to produce a measurable diffraction pattern [Arndt et al., 1999]. Matter-wave interference is not a special property of electrons; it is a universal feature of quantum objects, observable whenever a system can be prepared with a de Broglie wavelength comparable to some accessible length scale. The same phenomenon is the operating principle of the electron microscope, whose resolution — set by the wavelength of the imaging "light," per ordinary diffraction limits (Chapter 5) — can be far finer than any visible-light microscope because electron de Broglie wavelengths, as the worked example above shows, can be made far shorter than visible wavelengths simply by choosing a sufficiently large accelerating voltage.

Neutron Interferometry

Because neutrons are electrically neutral, they can be split into two coherent beams and recombined using **neutron interferometers** built from a single, precisely machined crystal of silicon, without the beam-steering complications that an electron's charge would introduce in a magnetic or electric field. A landmark 1975 experiment by Colella, Overhauser, and Werner (the "COW experiment") used exactly this technique to observe a measurable phase shift between the two arms of a neutron interferometer when one arm was raised slightly in the Earth's gravitational field relative to the other — a direct demonstration that a quantum matter wave, not merely a mathematical bookkeeping device, is affected by gravity precisely as its de Broglie wavelength and the classical gravitational potential energy predict [Colella et al., 1975]. Neutron interferometry has since been used to test the sign and magnitude of gravitational and even rotational (Sagnac) phase shifts on matter waves with high precision, extending the reach of wave–particle duality from crystal diffraction to macroscopic-scale sensitivity to gravity itself.

The Bohr Quantization Condition Revisited

The de Broglie hypothesis retroactively explains a feature of the 1913 Bohr model of the hydrogen atom (developed further in Chapter 10) that Bohr himself had simply postulated without justification: that an orbiting electron's angular momentum is quantized in integer multiples of $\hbar$, $L = n\hbar$. If the electron in a circular orbit of radius r is described by a de Broglie wave of wavelength $\lambda = h/p$, that wave can only form a consistent, single-valued standing pattern around the orbit if the orbit's circumference contains a whole number of wavelengths,

$$2\pi r = n\lambda = \frac{nh}{p}, \qquad n = 1, 2, 3, \ldots, \tag{7.5}$$

since otherwise the wave would interfere destructively with itself on successive trips around the loop and no stable pattern could persist. Rearranging, $pr = n\hbar$ — exactly Bohr's angular-momentum quantization condition, $L = n\hbar$, now derived (rather than assumed) from the

requirement that an electron's matter wave close consistently on itself. This does not yet constitute a full theory (that requires the Schrödinger equation of Chapters 8–9, applied to the hydrogen atom in Chapter 10), but it shows that de Broglie's hypothesis was not an isolated curiosity: it directly explains why atomic angular momentum comes only in discrete multiples of $\hbar$, years before Schrödinger's wave equation made the connection rigorous.

Quantization by a closure condition is not a quantum idea, and it is easier to trust once it has been seen somewhere unmysterious. Figure 7.4 drives a column of air and sweeps the frequency: almost every frequency produces nothing, and at a discrete set of them the tube suddenly resonates, because only there does the wave returning from the far end arrive back in step with itself. Nothing is quantized about air. What is discrete is the set of wavelengths that a boundary condition permits — an integer count fitting into the available length — and de Broglie's contribution was to notice that an electron's wave going around an orbit is subject to a closure condition of exactly the same kind.

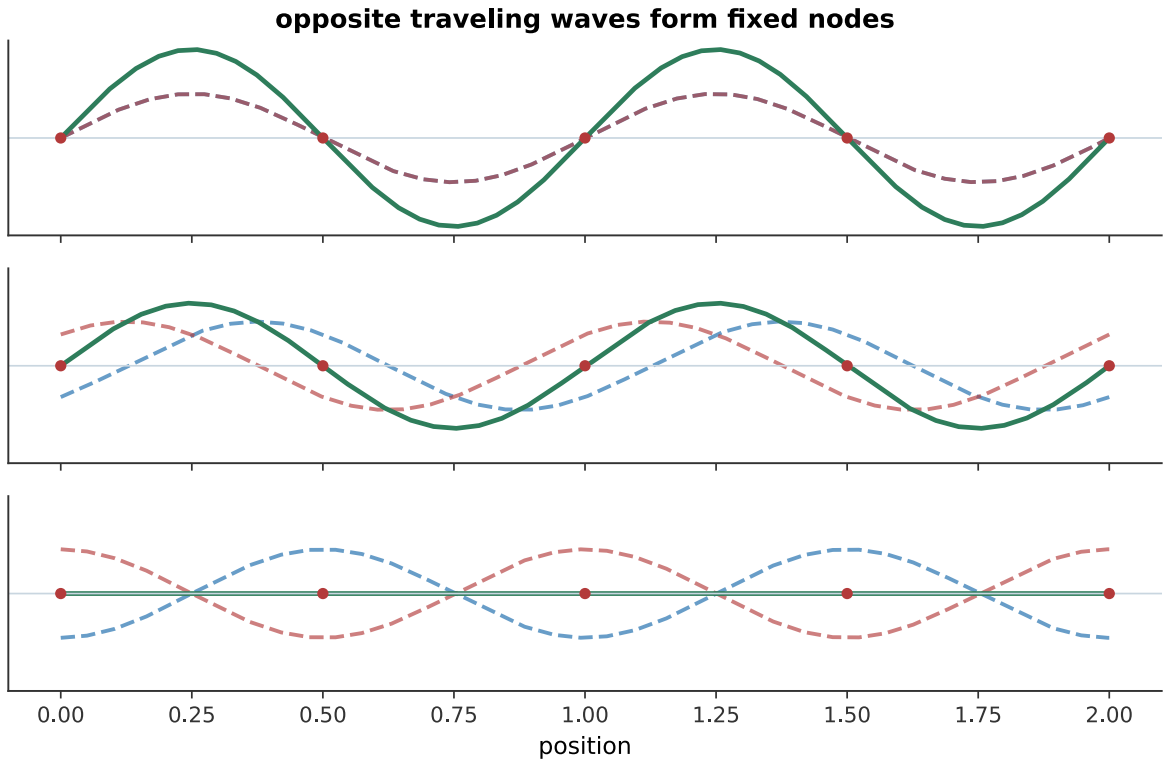

Figure 7.3: Two traveling waves moving in opposite directions add to a standing wave. The individual phases move, but the nodes remain fixed; boundary conditions can therefore select only wavelengths whose nodes land at the boundaries.

Worked Example: Checking the Standing-Wave Condition for Hydrogen's Ground State

The Bohr model (developed further in Chapter 10) gives the electron in hydrogen's $n = 1$ orbit a radius $a_0 = 5.29 \times 10^{-11}$ m (the **Bohr radius**) and an orbital speed $v = 2.19 \times 10^{6}$ m/s. Its de Broglie wavelength is

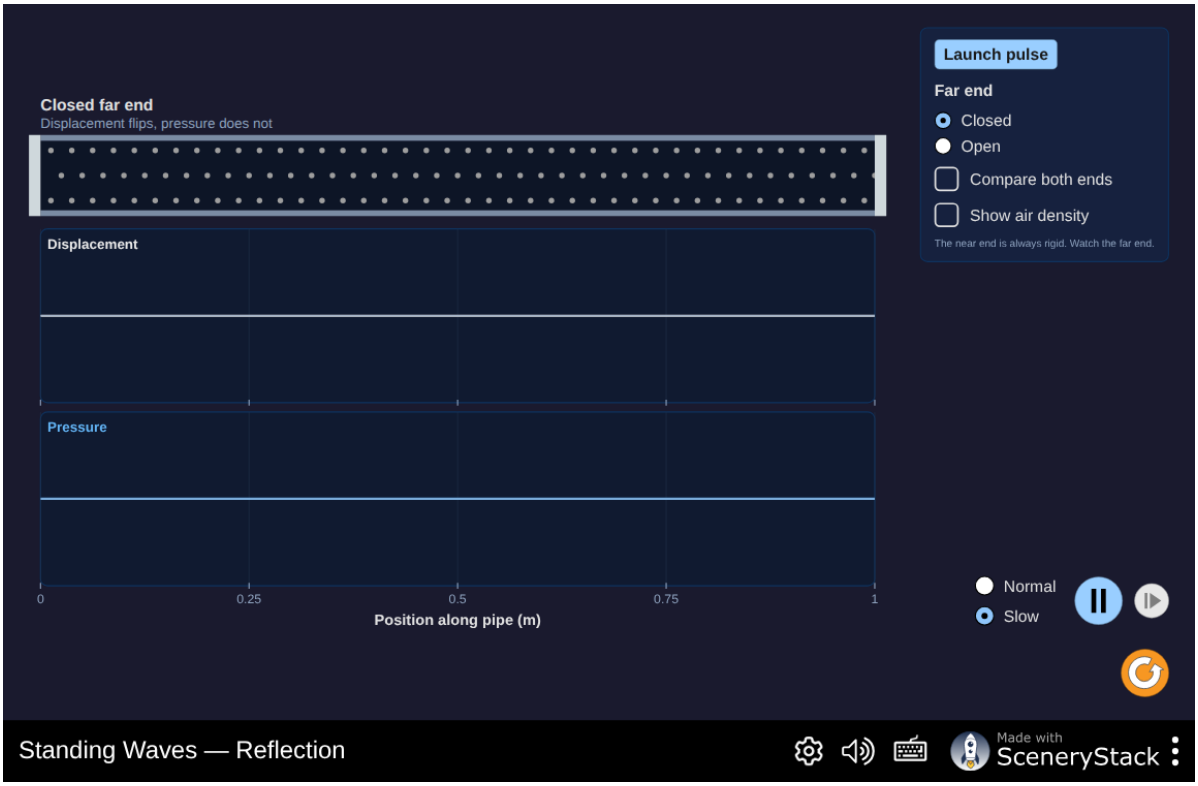

Figure 7.4: Standing waves in a pipe. Sweep the drive frequency and the response is a set of sharp resonances at which a whole number of half-wavelengths fits between the ends; between them, the returning wave arrives out of step and cancels itself. The Bohr condition $2\pi r = n\lambda$ is the same statement for a wave that closes on a loop instead of between two ends. Interactive simulation: Standing Waves

$$\lambda = \frac{h}{m_e v} = \frac{6.626 \times 10^{-34}\ \text{J}\cdot\text{s}}{(9.11 \times 10^{-31}\ \text{kg})(2.19 \times 10^{6}\ \text{m/s})} = 3.32 \times 10^{-10}\ \text{m}, \tag{7.6}$$

while the orbit's circumference is $2\pi a_0 = 2\pi(5.29 \times 10^{-11}\ \text{m}) = 3.32 \times 10^{-10}\ \text{m}$ — the two agree to three significant figures, exactly confirming the standing-wave condition $2\pi r = n\lambda$ for $n = 1$: the ground-state orbit's circumference is precisely one de Broglie wavelength.

7.2 Wave–Particle Duality and Wave Packets

Wave–Particle Duality

The picture that emerges from Chapters 6 and 7 together is symmetric: light, ordinarily described as a wave, exhibits particle-like behavior (photoelectric effect, Compton scattering); matter, ordinarily described as particles, exhibits wave-like behavior (electron diffraction). Neither description is simply "wrong" and replaced by the other; rather, **both light and matter possess both wave and particle aspects**, and which aspect is manifest depends on the experiment performed. This is **wave–particle duality**, and its sharpest expression is Niels Bohr's **principle of complementarity**: the wave and particle descriptions are both necessary for a complete account of quantum behavior, both cannot be exhibited in full simultaneously by the same experimental arrangement, and neither alone is sufficient.

The double-slit experiment (Chapter 4) makes this concrete. Even when electrons (or photons) are sent through the apparatus one at a time — far too infrequently for any two particles to interact or "interfere with each other" in transit — the individual, localized detection events recorded on a screen accumulate, after many particles, into the same interference fringe pattern predicted for a wave. Each particle arrives as a single, point-like click, exactly as a particle should; yet the *statistical distribution* of many such clicks builds up the wave-interference pattern, meaning each individual particle's behavior is governed by the same wave mathematics used for light, even though no two particles are present in the apparatus at once to "interfere" with each other in any classical sense. This experiment has been performed with electrons (originally by Claus Jönsson in 1961, and in a particularly clean single-particle-at-a-time form by Akira Tonomura and collaborators in 1989) and confirms the interference pattern exactly as quantum mechanics predicts.

That accumulation is the thing to watch, and it is what Figure 7.5 reproduces. Fire particles one at a time and the screen records isolated, point-like hits in what looks at first like a random scatter; leave it running and the fringes emerge from the statistics of hits that were never anything but individual. The simulation also carries the which-path apparatus discussed next: a detector placed in the path of the wave function collapses it on each measurement, and the fringes go with it, while the individual hits go on looking exactly the same.

Crucially, if a measurement is added to the apparatus that determines *which* slit each particle actually passed through — "which-path" information — the interference pattern disappears entirely, replaced by the simple sum of the two single-slit patterns, exactly as if the particles were classical objects going through one slit or the other. The loss of interference is not limited to a literal momentum kick: a which-path detector can also mark the alternatives by entangling the particle with another degree of freedom, such as polarization or an internal detector state. Once those alternatives are distinguishable even in principle, their amplitudes no longer combine to produce interference; an uncertainty-principle momentum disturbance is one important realization, not the general explanation. It is tempting, but incorrect, to imagine that a photon or electron is "really" a tiny wave packet that sometimes behaves like

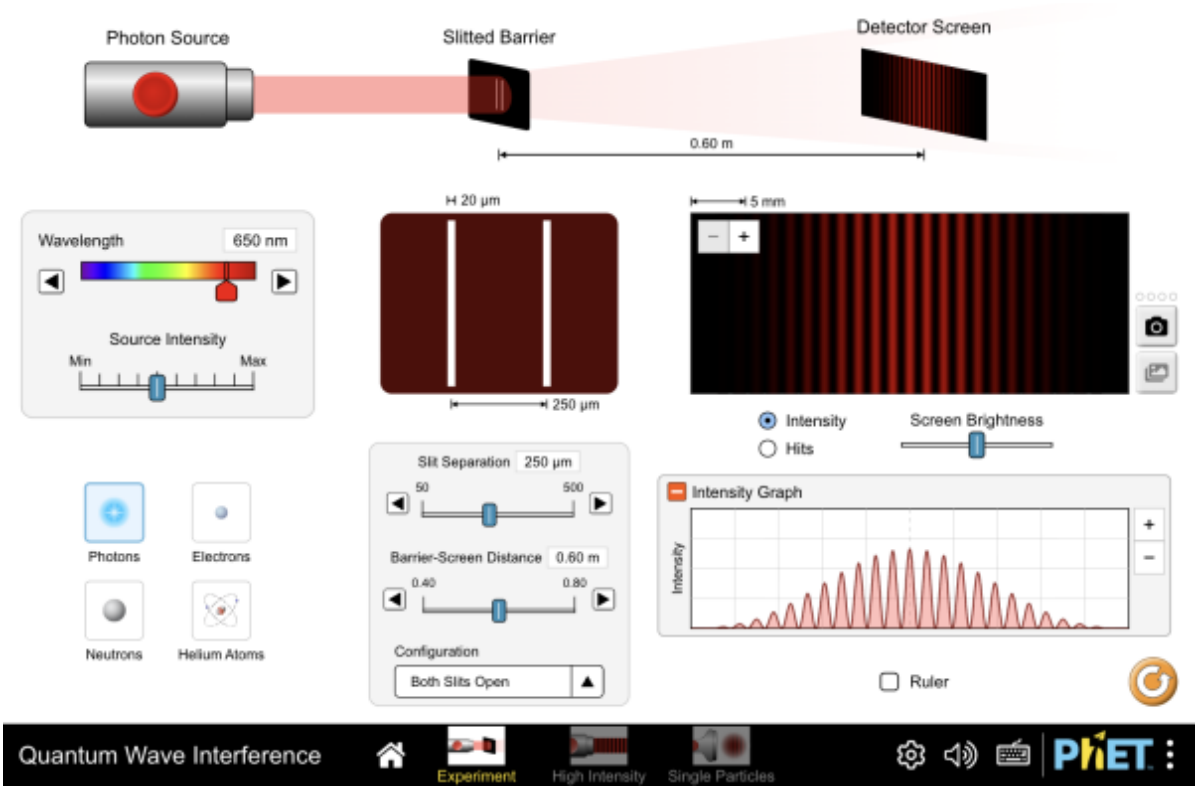

Figure 7.5: Single particles sent through a double slit, one at a time, with the detection screen accumulating hits. The wave function is displayed alongside the record of impacts: one object, propagating as a wave and detected as a particle. Interactive simulation: Quantum Wave Interference

a particle, or "really" a tiny particle that sometimes behaves like a wave; the two descriptions are complementary, and a full account of quantum behavior (developed starting in Chapter 8) requires a mathematical object — the wave function — that reduces to particle-like or wave-like predictions depending on what is measured, without being fully captured by either classical picture on its own.

Warning

The "measurement disturbs the momentum" language above is a useful first picture, but it is easy to over-read it as saying the uncertainty principle is merely a practical limitation of clumsy detectors that a sufficiently gentle measurement could someday evade. As the wave-packet discussion later in this chapter makes precise, the deeper reason is that a particle with a well-defined path through one specific slit cannot be described by a matter wave with a single well-defined wavelength; localizing the wave enough to carry which-path information *necessarily* broadens its spread in momentum. The trade-off lives in the mathematics of the wave itself, not in the clumsiness of any particular detector.

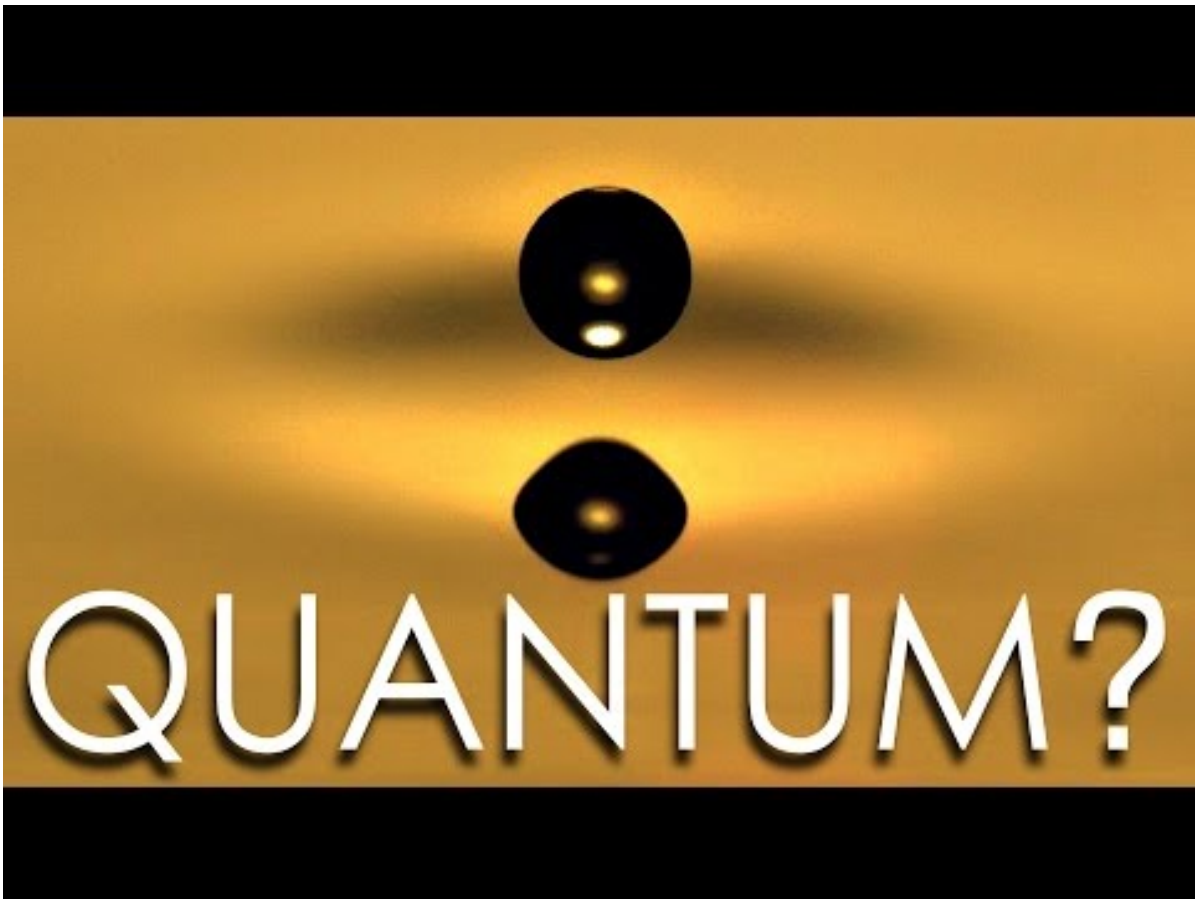

Figure 7.6: A macroscopic analogy — not a mechanism quantum mechanics itself uses — for how a single object might combine a definite, particle-like trajectory with a guiding wave, built from bouncing oil droplets rather than electrons. Watch on YouTube: Is This What Quantum Mechanics Looks Like?

See Also
Is Anything Real? (Vsauce) pushes the same question further: even the "solidity" that lets your hand feel a table is a quantum effect, ultimately traceable to the Pauli exclusion principle (Chapter 11) rather than to any classical picture of matter as tiny solid billiard balls.

Delayed Choice and the Quantum Eraser

A particularly striking variant, proposed by John Wheeler in 1978 and since realized experimentally, asks whether the decision to measure which-path information can be postponed until *after* a particle has, in some naive classical sense, "already passed" through the slits. In a **delayed-choice experiment**, the choice of whether to record which-path information (destroying the interference pattern) or to erase it before the particle is detected (restoring the interference pattern) is made only at the very last possible moment — in some realizations, only after the particle has already traversed the region of the slits. Experiments of this kind confirm quantum mechanics' prediction exactly: interference reappears in the appropriate measurement outcomes whenever the alternatives are made indistinguishable, regardless of when in the experiment that choice is made. A related arrangement, the **quantum eraser**, first correlates each particle with a "marker" that could reveal which path it took, then either reads that marker (destroying the interference pattern in the corresponding subset of detection events) or measures it in a basis that erases the path distinction (revealing interference only in appropriately sorted subsets). Neither experiment allows sending a signal backward in time or violates causality. If the marker result is ignored, the unconditional detection pattern remains interference-free; the fringes in a sorted subset become visible only after the corresponding marker records are compared. These results reinforce the same lesson as the ordinary double-slit experiment: a quantum system does not carry a hidden, predetermined trajectory waiting to be revealed, and complementarity is a statement about what information can coexist, not merely about the limits of measurement technology.

Wave Packets

To describe a localized particle in wave language, a single wave of definite wavelength $\lambda = h/p$ — which by its nature extends infinitely in space with constant amplitude — is not adequate, since it corresponds to a perfectly definite momentum but gives no information about *where* the particle is. A localized particle is instead represented by a **wave packet**: a superposition of many waves of slightly different wavelength (equivalently, different wave number $k = 2\pi/\lambda$), chosen so that they interfere constructively in some limited region Δx and destructively (cancel) elsewhere. The mathematics of superposition (the same mathematics used for beats and Fourier synthesis of waveforms) requires that a packet localized to a narrow spatial region Δx necessarily be built from a *broad* range of wave numbers Δk, and vice versa; the two spreads are inversely related, roughly

$$\Delta x \, \Delta k \gtrsim 1, \tag{7.7}$$

a purely mathematical fact about waves, true for sound pulses and water-wave packets just as much as for matter waves, with no quantum content yet.

The trade-off can be watched directly. Figure 7.9 builds a packet out of harmonic components whose amplitudes follow an adjustable envelope, and prints both widths as you work:

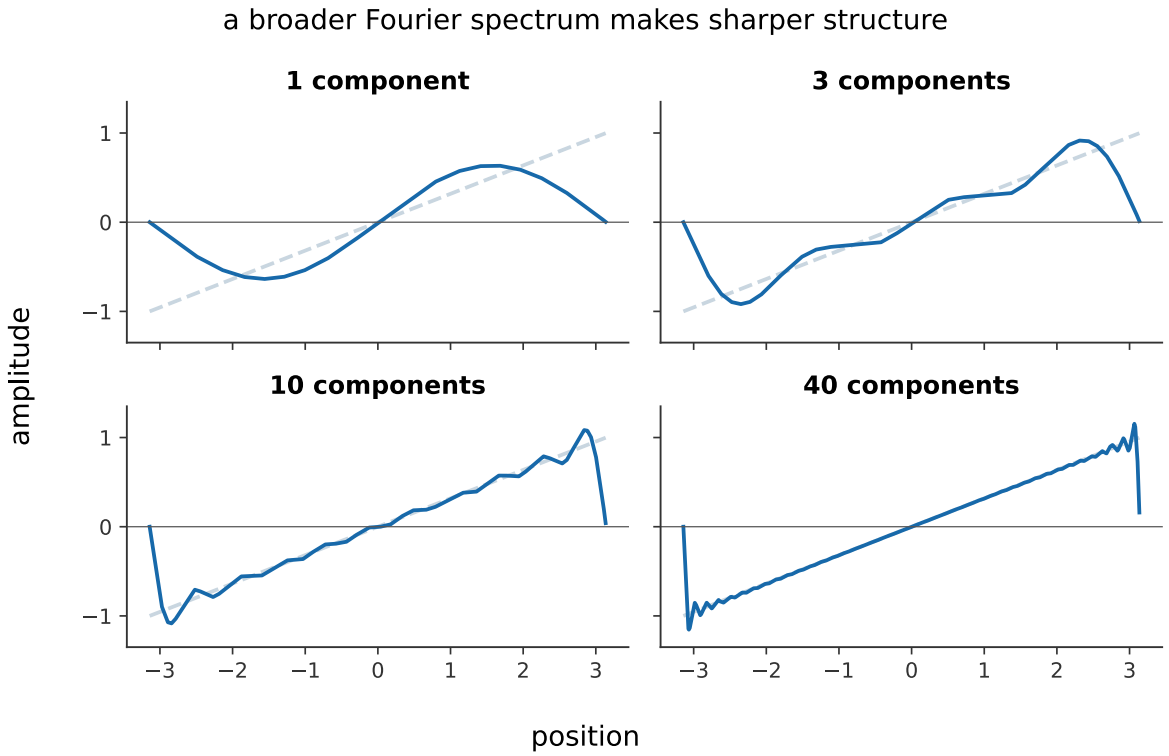

Figure 7.7: Fourier synthesis in a deliberately non-quantum setting. A single sinusoid is spread smoothly across the whole interval; adding a wider range of wave numbers builds sharper spatial structure. A wave packet uses the same mathematical mechanism, with coefficients chosen to make one localized pulse rather than a repeating sawtooth.

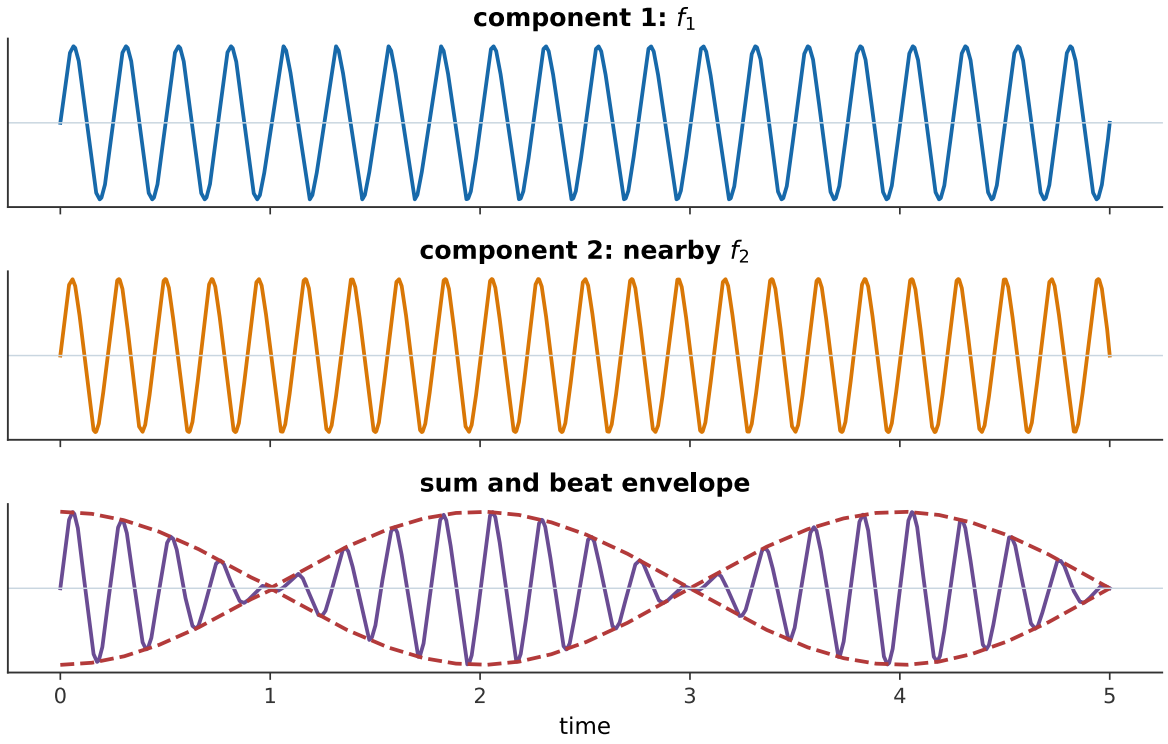

Figure 7.8: The smallest possible packet uses only two nearby frequencies. Their rapidly oscillating sum is carried by a slow beat envelope, the elementary version of the group-velocity construction developed below.

narrow the spread σ_k of contributing wave numbers and the packet in x stretches out toward the infinite sinusoid of definite momentum; widen σ_k and the packet contracts toward a spike, with the panel reporting $\sigma_x = 1/\sigma_k$ throughout. Switching the display from a function of space to a function of time turns the same relation into the energy–time form used later in this section.

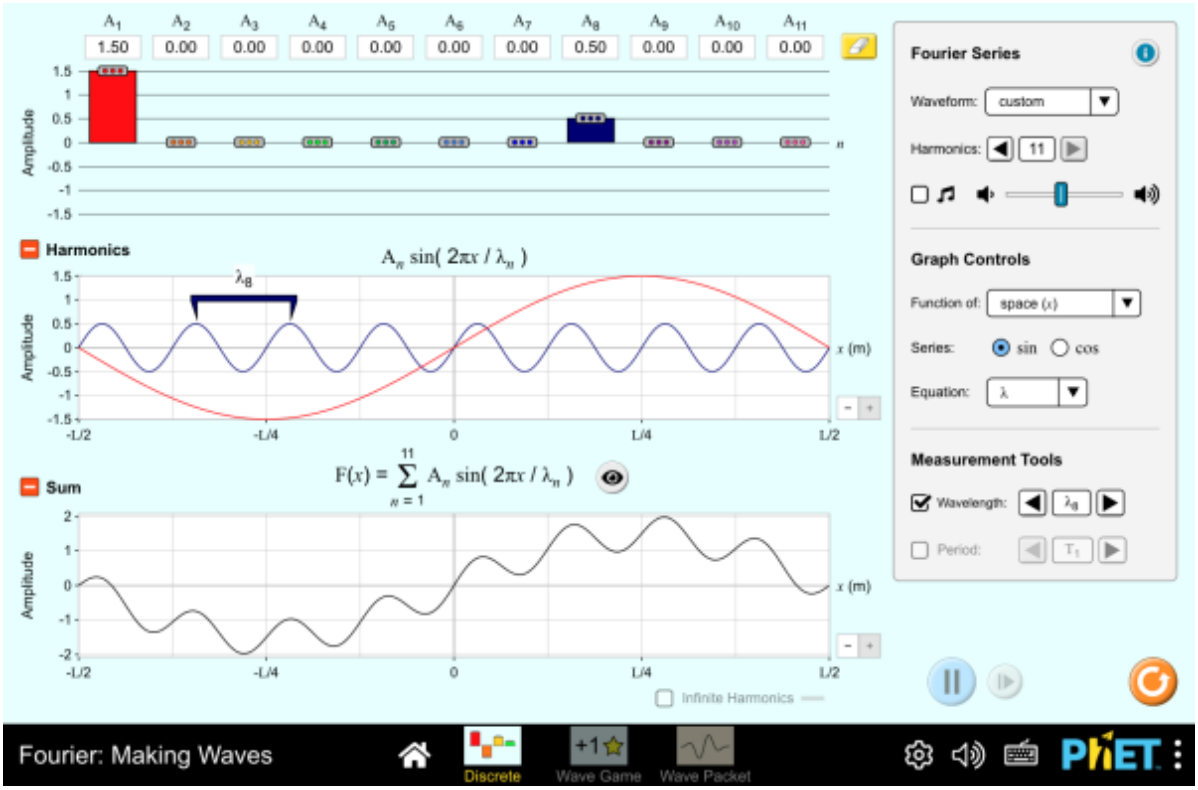

Figure 7.9: A wave packet and the Fourier components that build it, with the width in k and the width in x both displayed. Their product cannot be reduced: this is the classical wave theorem that becomes the uncertainty principle once $p = \hbar k$ is imposed. Interactive simulation: Fourier: Making Waves

Quantum mechanics enters when this relation is combined with the de Broglie relation $p = hk/2\pi = \hbar k$ (where $\hbar \equiv h/2\pi$), converting spread in wave number into spread in momentum, $\Delta p = \hbar\,\Delta k$, and giving

$$\Delta x\,\Delta p \gtrsim \hbar. \tag{7.8}$$

Phase Velocity and Group Velocity

A wave packet built from components of angular frequency $\omega(k)$ spread over a narrow range of k can be shown, by adding two nearby component waves $\cos(kx - \omega t)$ and $\cos[(k + \Delta k)x - (\omega + \Delta\omega)t]$, to have an overall envelope that moves at the **group velocity**

$$v_g = \frac{d\omega}{dk}, \tag{7.9}$$

which is, in general, different from the **phase velocity** $v_p = \omega/k$ at which the individual wave crests inside the envelope move.

A mnemonic: the **group** velocity is the speed of the **group** — the envelope, the localized "blob" that carries the particle and any information — while the **phase** velocity is the speed of the individual **phase** fronts (the crests) sliding along underneath that envelope. Only the first one is a particle speed.

For a nonrelativistic free particle, the de Broglie relations $p = \hbar k$ and $E = \hbar\omega$, combined with the nonrelativistic kinetic-energy relation $E = p^2/2m$, give

$$\omega(k) = \frac{\hbar k^2}{2m}, \tag{7.10}$$

so that

$$v_g = \frac{d\omega}{dk} = \frac{\hbar k}{m} = \frac{p}{m} = u, \tag{7.11}$$

exactly the classical particle velocity — the wave packet's envelope, the physically observable, localized "blob" of probability, moves at precisely the speed a classical particle with the same momentum would have. The phase velocity, by contrast, is $v_p = \omega/k = \hbar k/2m = u/2$, exactly *half* the particle's actual speed: the individual crests inside the packet move at a different, less physically meaningful speed than the envelope itself, a reminder that it is the group velocity, not the phase velocity, that corresponds to the motion of the particle (and, for a light pulse in an ordinary, weakly dispersive medium, that closely tracks the speed at which energy and information travel — the two velocities can part ways near a strong absorption line, but that subtlety does not arise here).

Deriving the envelope speed from two superposed waves

The group-velocity formula $v_g = d\omega/dk$ can be seen directly by adding just two component waves of equal amplitude, wave numbers $k \pm \Delta k/2$, and angular frequencies $\omega \pm \Delta\omega/2$:

$$\cos\left[\left(k + \tfrac{\Delta k}{2}\right) x - \left(\omega + \tfrac{\Delta\omega}{2}\right) t\right] + \cos\left[\left(k - \tfrac{\Delta k}{2}\right) x - \left(\omega - \tfrac{\Delta\omega}{2}\right) t\right]. \tag{7.12}$$

Using the sum-to-product identity $\cos A + \cos B = 2\cos\left(\frac{A+B}{2}\right)\cos\left(\frac{A-B}{2}\right)$ with A and B the two bracketed phases gives

$$2\cos(kx - \omega t)\,\cos\left(\frac{\Delta k}{2}x - \frac{\Delta\omega}{2}t\right). \tag{7.13}$$

The first factor is a rapidly oscillating carrier wave moving at the phase velocity ω/k; the second factor is a slowly varying envelope — the "beat" pattern familiar from superposing two close sound frequencies — whose crests (where the cosine argument is constant) move at speed $\Delta\omega/\Delta k$. Adding more and more component waves in a narrow band around k sharpens this beat pattern into a single, well-localized wave packet, and in the limit $\Delta k \to 0$ the envelope speed $\Delta\omega/\Delta k$ becomes exactly $d\omega/dk$, the group velocity used in the main text.

Wave Packet Spreading

The nonrelativistic dispersion relation in Equation (7.10) is not linear in k (unlike the dispersion relation $\omega = ck$ for light in vacuum), which has a further consequence beyond fixing the group velocity: different Fourier components of a wave packet, corresponding to different momenta, travel at slightly different group velocities $v_g(k) = \hbar k/m$, since a component with larger k (larger momentum) simply moves faster. A packet initially localized to a narrow width Δx_0

therefore does not maintain its shape as it propagates; because it necessarily contains a spread $\Delta k \sim 1/\Delta x_0$ of wave numbers (the same uncertainty relation used earlier in this section), and each component travels at a slightly different speed, the packet's spatial width grows with time — it **spreads**. A careful calculation shows the spreading time scale is of order $\tau \sim m(\Delta x_0)^2/\hbar$: a light particle (small m) confined to a very small initial region (small Δx_0) spreads apart quickly, while a macroscopic object's wave packet spreads on a time scale so astronomically long that the effect is entirely unobservable — one more reason, alongside the negligible de Broglie wavelength itself, that ordinary macroscopic objects appear to follow sharp, well-defined classical trajectories rather than visibly diffusing.

7.3 The Heisenberg Uncertainty Principle

Werner Heisenberg (1927) elevated the wave-packet relation above to a fundamental principle governing all quantum systems [Heisenberg, 1927], stated precisely as

$$\Delta x \, \Delta p_x \geq \frac{\hbar}{2}, \tag{7.14}$$

where Δx and Δp_x are, more precisely, statistical spreads (standard deviations) in simultaneous measurements of position and momentum made on identically prepared systems. The **Heisenberg uncertainty principle** states that these two spreads cannot both be made arbitrarily small: the more precisely a particle's position is known, the less precisely its momentum can be known, and conversely. This is not a statement about the clumsiness of measuring instruments, correctable in principle by better technology — it is a consequence of the wave nature of matter itself, as the wave-packet argument above shows: a particle simply *does not possess* simultaneously well-defined position and momentum, in the same sense that a wave pulse of well-defined wavelength cannot also be localized to a point.

> **See Also**
> The Schrödinger Equation puts this wave-packet picture on rigorous footing: the wave function $\Psi(x,t)$ introduced there replaces the qualitative superposition argument of this chapter with an exact quantum-mechanical object, from which $\Delta x \, \Delta p_x \geq \hbar/2$ can be derived precisely rather than estimated.

An analogous relation holds between energy and time,

$$\Delta E \, \Delta t \geq \frac{\hbar}{2}, \tag{7.15}$$

where Δt characterizes the time available to measure (or the lifetime of a state with) energy spread ΔE. This relation, for instance, explains why an unstable state with a short lifetime Δt (such as an excited atomic state, or an unstable particle) necessarily has an intrinsic spread, or "width," in its energy — and correspondingly in the frequency/wavelength of radiation it emits — that grows as its lifetime shrinks.

Stripped of $\hbar$, this is a theorem about signals rather than about quantum mechanics, and it applies to any wave one cares to measure — including the sound of one's own voice. Figure 7.11 runs a live spectrum of whatever the microphone hears: a sustained vowel, going on for a long Δt, resolves into sharp harmonic lines with small Δf, while a clipped consonant lasting a few milliseconds has no sharp lines at all, only a broad smear across the spectrum. No better

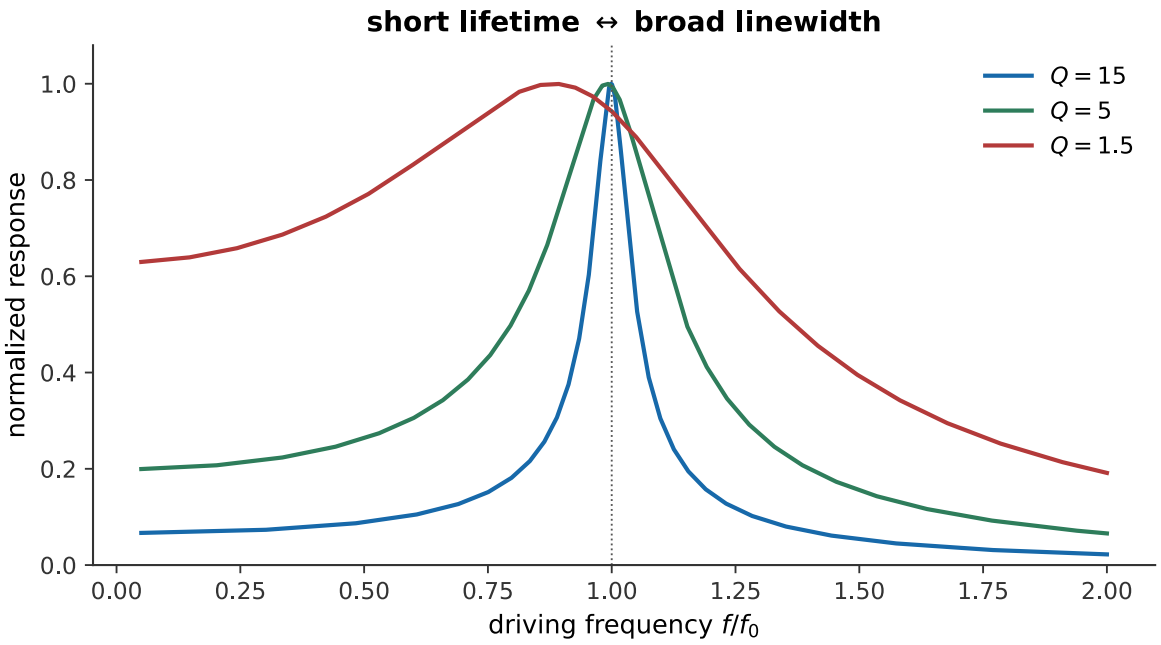

Figure 7.10: Resonance makes linewidth visible. Long-lived, weakly damped oscillations have narrow frequency peaks; short-lived, strongly damped oscillations have broad peaks. Atomic line widths and unstable-particle decay widths express the same lifetime–bandwidth connection.

microphone would fix this. A short signal does not *have* a well-defined frequency, exactly as a short-lived excited state does not have a well-defined energy.

Worked Example: Confining an Electron in a Nucleus

Could an electron exist bound inside an atomic nucleus, of radius $r \sim 5 \times 10^{-15}$ m? If so, the position uncertainty could be no larger than $\Delta x \sim r$, and the uncertainty principle then requires a momentum uncertainty of at least

$$\Delta p \gtrsim \frac{\hbar}{2\,\Delta x} \approx \frac{1.055 \times 10^{-34}\ \text{J}\cdot\text{s}}{2(5 \times 10^{-15}\ \text{m})} \approx 1.1 \times 10^{-20}\ \text{kg}\cdot\text{m/s}. \tag{7.16}$$

Converting to an energy via the (relativistic, since this momentum turns out to be large) relation $E \approx pc$ for $pc \gg mc^2$: $pc \approx (1.1 \times 10^{-20}\ \text{kg}\cdot\text{m/s})(3.0 \times 10^{8}\ \text{m/s}) \approx 3.3 \times 10^{-12}$ J $\approx$ 21 MeV. An electron confined to nuclear dimensions would need kinetic energy of tens of MeV — far larger than the few-MeV binding energies available in nuclei (Chapter 13) — so such an electron could not remain bound; this is one of the historical arguments (alongside others involving nuclear spin and magnetic moment) that electrons are not constituents of the nucleus, correctly anticipating that beta decay (Chapter 13) must *create* an electron at the moment of decay rather than releasing one that was previously confined inside.

Worked Example: The Scale of Atomic Energies

The uncertainty principle also correctly predicts the *order of magnitude* of atomic binding energies, without needing to solve the Schrödinger equation at all. An electron confined to an

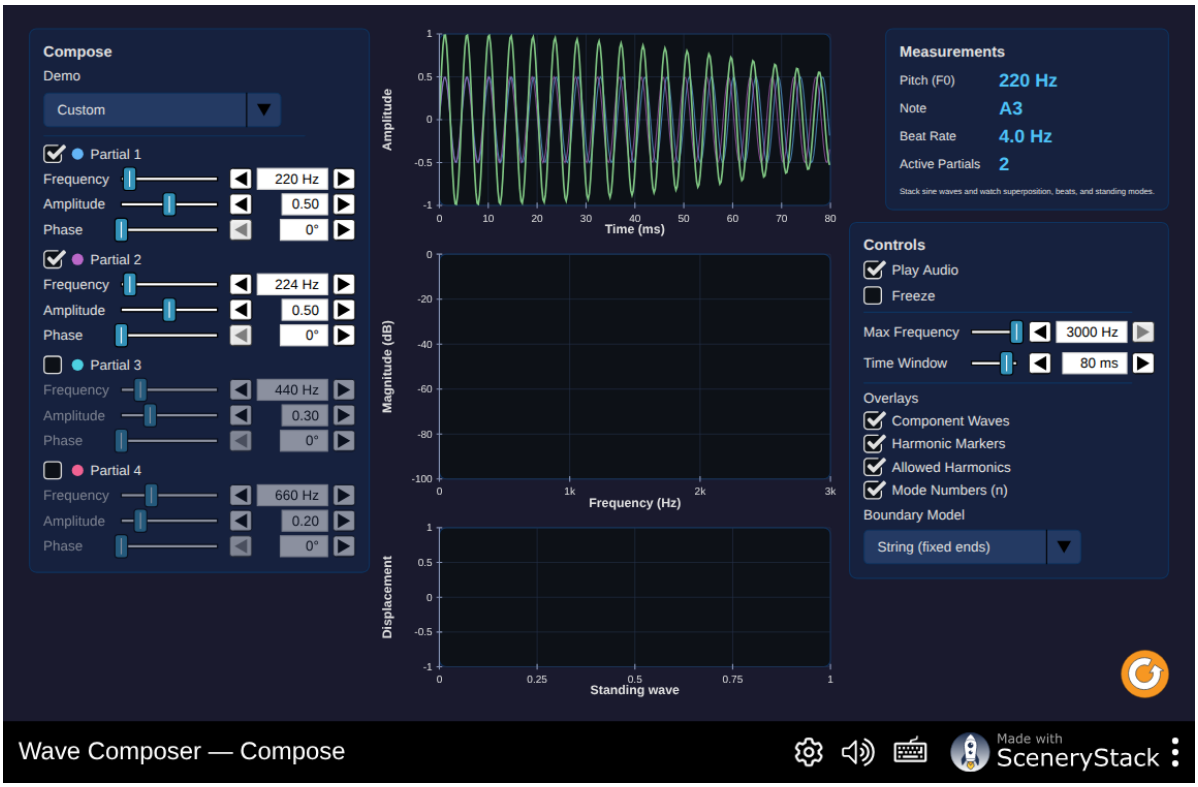

Figure 7.11: Real-time spectrum analysis of a microphone signal. The duration of a sound and the sharpness of its spectrum trade off against each other, which is $\Delta E\,\Delta t \gtrsim \hbar/2$ with the $\hbar$ removed and the quantum mechanics along with it. Interactive simulation: Wave Composer

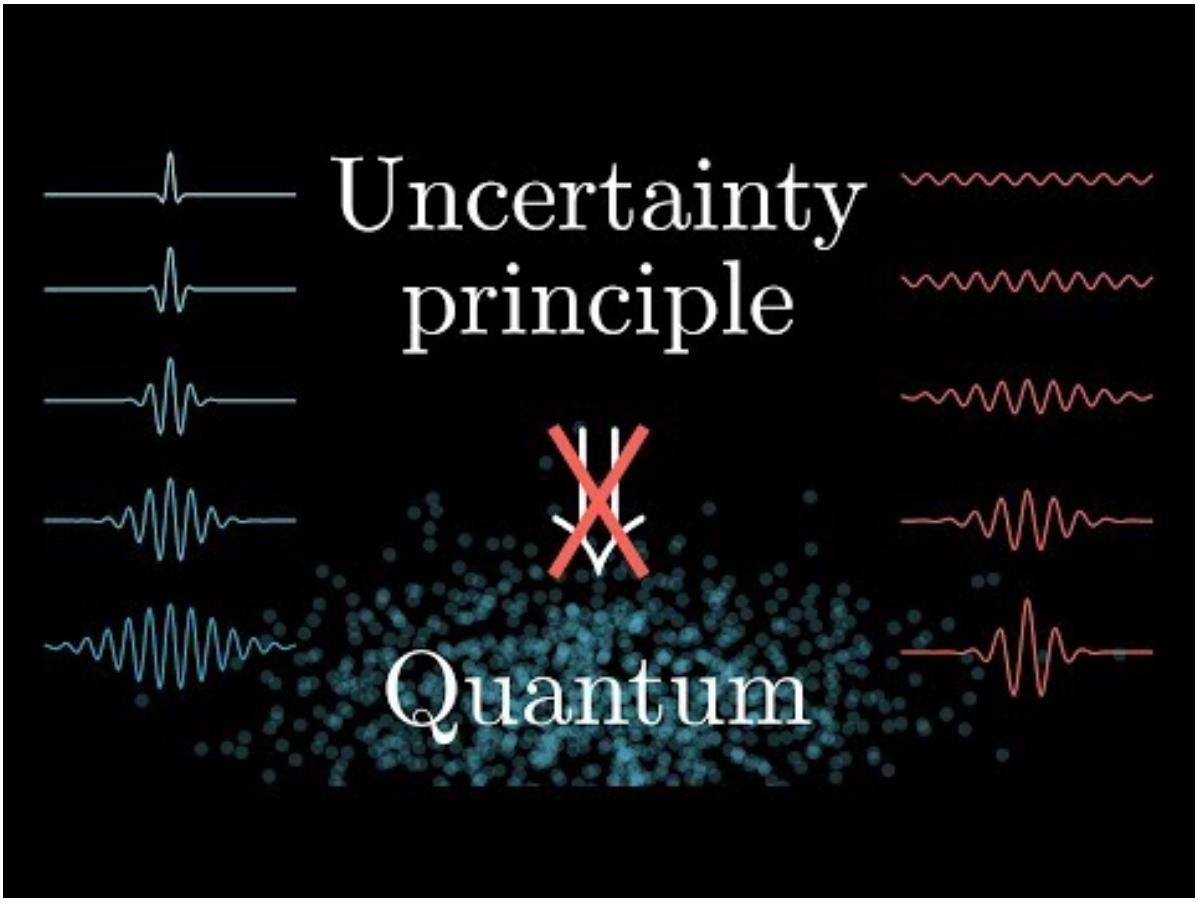

Figure 7.12: A rigorous version of the "no better microphone would fix this" argument above: $\Delta x\,\Delta p_x \geq \hbar/2$ as one instance of a general Fourier-transform trade-off between any conjugate pair of variables. Watch on YouTube: The more general uncertainty principle, regarding Fourier transforms



Figure 7.13: A sense of scale for the numbers in the worked examples that follow: how far down the uncertainty principle and quantum mechanics actually reach before a different kind of physics takes over. Watch on YouTube: What Is The Smallest Possible Thing?

atom of radius $r \sim a_0 = 5.3 \times 10^{-11}$ m (the Bohr radius, derived properly in Chapter 10) has, by the same reasoning as above, a minimum momentum uncertainty

$$\Delta p \gtrsim \frac{\hbar}{2\,\Delta x} \approx \frac{1.055 \times 10^{-34}\ \text{J}\cdot\text{s}}{2(5.3 \times 10^{-11}\ \text{m})} \approx 1.0 \times 10^{-24}\ \text{kg}\cdot\text{m/s}. \tag{7.17}$$

Since this is small enough that the electron remains nonrelativistic, its kinetic energy is $K \sim (\Delta p)^2/2m_e \approx (1.0 \times 10^{-24}\ \text{kg}\cdot\text{m/s})^2/[2(9.11 \times 10^{-31}\ \text{kg})] \approx 5.5 \times 10^{-19}$ J ≈ 3.4 eV — the right order of magnitude for atomic binding and ionization energies (a few to a few tens of eV), even though the precise numerical value requires the full machinery of Chapters 9–10. This is a recurring pattern in quantum mechanics: the uncertainty principle alone, applied as an order-of-magnitude estimate, correctly anticipates the characteristic energy scale of a confined system, well before an exact calculation is carried out.

7.4 Summary

- The **de Broglie hypothesis** assigns every particle of momentum p a wavelength $\lambda = h/p$, extending the photon relation of Chapter 6 to all matter, and for an electron accelerated through voltage V gives $\lambda \approx 1.226\ \text{nm}/\sqrt{V[\text{volts}]}$.
- Requiring an electron's de Broglie wave to form a consistent standing pattern around a circular orbit, $2\pi r = n\lambda$, reproduces the Bohr model's angular-momentum quantization $L = n\hbar$, previously an unexplained postulate.
- The **Davisson–Germer** and **G. P. Thomson** experiments confirmed matter waves directly: electrons diffract from a crystal lattice with a wavelength matching $\lambda = h/p$, exactly as X-rays do; matter-wave interference has since been observed for neutrons, atoms, and large molecules such as C_{60}.
- **Wave–particle duality** and Bohr's **complementarity principle**: light and matter both show wave and particle behavior; which is manifest depends on the experiment (as the single-particle double-slit experiment and the loss of fringes upon acquiring which-path information both show), and neither classical picture alone is complete.
- A localized particle is represented by a **wave packet**, a superposition of waves over a range of wave numbers Δk; its envelope moves at the **group velocity** $v_g = d\omega/dk$, which equals the particle's classical velocity, distinct from the **phase velocity** $v_p = \omega/k$ of the individual wave crests.
- The **Heisenberg uncertainty principle**, $\Delta x\,\Delta p_x \geq \hbar/2$ (and analogously $\Delta E\,\Delta t \geq \hbar/2$), is a fundamental limit on the simultaneous precision of conjugate quantities, rooted in the wave nature of matter, not a limitation of measuring instruments, and correctly predicts the order of magnitude of confinement energies from nuclear to atomic scales.

7.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens on a wave-packet envelope.

1. **Multiple choice, from a figure.** How does doubling a particle's momentum change its de Broglie wavelength? Figure description: Two nearby matter-wave components and the envelope produced by their superposition. Choices: (A) It halves the wavelength. (B) It

doubles the wavelength. (C) It quadruples the wavelength. (D) It leaves the wavelength unchanged.

2. **True or false.** For a free nonrelativistic particle, the group velocity of its wave packet equals the particle speed.

3. **Drag the words.** Complete the matter-wave relationships. Matter wavelength is _____. A localized particle requires a _____. Its envelope moves at the _____, while its individual crests move at the _____. (Terms: $\lambda = h/p$; wave packet; group velocity; phase velocity; $\lambda = p/h$; single plane wave.)

4. **Fill in the blanks.** Complete the wavelength and uncertainty comparisons. Doubling momentum makes λ _____ as large; narrowing Δx requires the momentum spread Δp to become _____.

5. **Mark the words.** Mark the two experimental signatures of matter waves. "Electrons produce crystal diffraction peaks and individual particles build up a two-path interference pattern. A localized detector click is the particle-like part of the same experiment."

7.6 Conceptual Questions

1. Why is electron diffraction evidence for a wave property rather than merely complicated particle scattering?

2. How does the standing-wave interpretation of the Bohr condition replace an arbitrary rule with a boundary condition?

3. Why does obtaining which-path information destroy a two-path interference pattern even if the detector does not appreciably push the particle?

4. A broad matter wave has a sharply defined momentum. Explain why it must be poorly localized.

5. Distinguish phase velocity from group velocity and identify which can carry a localized signal.

6. Why does a short-lived state have an intrinsic energy width even with a perfect spectrometer?

7.7 Problems

Exercise 7.1
(Straightforward)

Find the de Broglie wavelength of (a) an electron with kinetic energy 54 eV (as in the original Davisson–Germer experiment), (b) a proton with kinetic energy 1.0 MeV, and (c) a 0.145 kg baseball moving at 40 m/s. Comment on which of these wavelengths could plausibly produce observable diffraction, and from what kind of structure.

Exercise 7.2

(Straightforward)

In the Davisson–Germer experiment, a diffraction maximum for 54 eV electrons was observed at $\theta = 50°$ from a nickel crystal. Using $\lambda = h/p$ for the electrons' de Broglie wavelength and the Bragg condition $2d\sin\theta = n\lambda$ (first order, $n = 1$), find the effective interplanar spacing d of the nickel crystal consistent with this observation.

Exercise 7.3

(Straightforward)

Using the worked-example formula $\lambda \approx 1.226\ \text{nm}/\sqrt{V[\text{volts}]}$, find the accelerating voltage needed to give electrons a de Broglie wavelength of exactly 0.0500 nm, comparable to a typical X-ray wavelength used in crystallography.

Exercise 7.4

(Straightforward)

A beam of thermal neutrons (kinetic energy $\approx k_B T$ at room temperature, $T \approx 300$ K) is used for neutron diffraction studies of crystal structure. (a) Estimate the neutrons' typical kinetic energy in eV. (b) Find their de Broglie wavelength, given the neutron mass $m_n = 1.675 \times 10^{-27}$ kg, and compare it to typical interatomic spacings.

Exercise 7.5

(Straightforward)

Verify the claim in the text that a C_{60} molecule (mass $\approx 1.2 \times 10^{-24}$ kg) moving at a typical thermal beam speed of 200 m/s has a de Broglie wavelength of a few picometers, and comment on why such a short wavelength, far smaller than the molecule's own diameter (≈ 1 nm), can nonetheless produce an observable diffraction pattern from a grating with a period of order 100 nm.

Exercise 7.6

(Straightforward)

Using the de Broglie standing-wave argument, show that an electron in a circular Bohr orbit with $n = 1$ has an orbital circumference equal to exactly one de Broglie wavelength, and explain qualitatively (without calculation) why this is consistent with such an orbit being the atom's lowest-energy (ground) state.

Exercise 7.7

(Straightforward)

A proton is confined to a nucleus of diameter 1.0×10^{-14} m. Use the uncertainty principle to estimate the minimum kinetic energy the proton must have, and compare it (order of magnitude) to typical nuclear binding energies of several MeV per nucleon (Chapter 13).

Exercise 7.8

(Straightforward)

An excited atomic state has a mean lifetime of $\Delta t = 1.0 \times 10^{-8}$ s. (a) Use the energy–time uncertainty relation to estimate the energy spread ΔE of this state, in eV. (b) If the state decays by emitting a photon of wavelength 500 nm, estimate the corresponding spread (linewidth) $\Delta\lambda$ in the emitted wavelength.

Exercise 7.9
(Moderate)

A beam of electrons is passed through a single slit of width a. Using the single-slit diffraction condition (first minimum at $\sin\theta \approx \lambda/a$ for small angles) together with the de Broglie relation, express the transverse momentum spread Δp_y imparted to the electrons (estimated from $p\sin\theta$) in terms of the slit width a, and show that $\Delta y\,\Delta p_y \sim h$ if $\Delta y \sim a$, consistent with the uncertainty principle.

Exercise 7.10
(Moderate)

Explain, using the energy–time uncertainty relation, why a particle that is truly stable (infinite lifetime) can have a perfectly sharp rest energy mc^2, while an unstable particle cannot — and why particle physicists therefore quote both a mass and a "width" (in energy units) for unstable particles, a topic revisited in Chapter 14.

Exercise 7.11
(Moderate)

Verify the group-velocity calculation in the text: starting from $\omega(k) = \hbar k^2/2m$ (the nonrelativistic free-particle dispersion relation), compute $d\omega/dk$ explicitly and confirm that it equals $\hbar k/m = p/m$.

Exercise 7.12
(Moderate)

A relativistic free particle has energy $E = \sqrt{(pc)^2 + (mc^2)^2}$ (Chapter 3). Using $E = \hbar\omega$ and $p = \hbar k$, show that the group velocity $v_g = d\omega/dk$ equals pc^2/E, and confirm that this reduces to the particle's actual (relativistic) velocity u using the relation $u = pc^2/E$ established in Chapter 3.

Exercise 7.13
(Moderate)

Estimate, using the uncertainty principle in the same manner as the atomic-scale worked example, the characteristic kinetic energy of an electron confined to a quantum dot (a nanoscale semiconductor structure) of diameter 10 nm, and compare it in order of magnitude to the atomic-scale estimate found in the text.

Exercise 7.14
(Moderate)

A double-slit experiment is performed with electrons at a rate of one electron at a time, with a detector placed at one slit to determine which slit each electron passes through. (a) Explain, in terms of the uncertainty principle, why this which-path measurement necessarily disturbs the electron's momentum. (b) Estimate the momentum disturbance needed to determine, to within the slit width a, which of two slits separated by distance d an electron passed through, and explain qualitatively why a disturbance of this size is enough to wash out an interference pattern with fringe spacing set by d.

Exercise 7.15
(Moderate)

Explain why the phase velocity $v_p = \omega/k$ of a nonrelativistic free-particle matter wave, found in the text to be $u/2$, is not the speed of anything physically observable (in contrast with the group velocity), and why this is not a paradox.

Exercise 7.16
(Moderate)

Using the spreading time scale $\tau \sim m(\Delta x_0)^2/\hbar$ quoted in the text, estimate τ for (a) an electron initially localized to $\Delta x_0 = 1.0 \times 10^{-10}$ m (roughly an atomic diameter), and (b) a 1.0 mg dust grain initially localized to $\Delta x_0 = 1.0 \times 10^{-6}$ m. Comment on which of these spreading times could plausibly be observed in a laboratory measurement lasting a few seconds.

Exercise 7.17
(Challenging)

In a delayed-choice or quantum-eraser experiment, explain why the disappearance and reappearance of the interference pattern cannot be used to send information backward in time, even though the "choice" of whether which-path information is available may be made after the particle has passed through the slit region. (Hint: consider what an experimenter examining only the particles that land in a single, fixed detector region can and cannot know without also consulting the marker/eraser results.)

Exercise 7.18
(Challenging)

A beam of electrons, each with the same de Broglie wavelength $\lambda = h/p$, is used in a Davisson–Germer-type experiment. Explain, using the standing-wave condition of the "Bohr Quantization Condition Revisited" subsection as an analogy, why a diffraction condition $2d\sin\theta = n\lambda$ (rather than an arbitrary relation between θ and λ) is required for a diffraction maximum, drawing the parallel between constructive interference around a crystal lattice and constructive interference around a closed atomic orbit.

Exercise 7.19
(Challenging)

Verify the claim in the "Consistency Check" subsection that $E = pc^2/u$ for a relativistic particle, using $E = \gamma mc^2$ and $p = \gamma mu$ from Chapter 3, and show that combining this with $E = hf$ and $\lambda = v_p/f$ (with v_p the wave's phase velocity) gives a phase velocity $v_p = c^2/u$ greater than c — then explain why this superluminal *phase* velocity does not violate relativity, referring to your answer to Problem 15 about which velocity is physically observable.

Exercise 7.20
(Challenging)

The neutron interferometry (COW) experiment detects a gravitationally induced phase shift between two paths of different height Δh in a neutron interferometer of horizontal path length L. Explain qualitatively, using the de Broglie relation $\lambda = h/p$ and the fact that a neutron's kinetic energy (and hence its momentum and wavelength) changes very slightly with height in a gravitational field, why raising one arm of the interferometer changes the phase accumulated along that arm relative to the other.

Chapter 8

The Schrödinger Equation

Learning Objectives

By the end of this chapter, you should be able to:

- State the time-dependent and time-independent Schrödinger equations and explain the role of each.
- Interpret the wave function $\Psi(x, t)$ statistically (the Born interpretation) and apply normalization.
- State and apply the boundary and continuity conditions on acceptable wave functions.
- Compute expectation values of position and momentum from a wave function using the position and momentum operators.
- Solve the time-independent Schrödinger equation for a particle in an infinite square well and interpret the resulting quantized energies and stationary states.
- Analyze a potential step and compute reflection and transmission coefficients, and explain why quantum mechanics allows partial reflection even when $E > V_0$.
- Solve (qualitatively and, for simple cases, quantitatively) the finite square well and explain barrier penetration and quantum tunneling.
- Solve the quantum harmonic oscillator, apply its selection rule, and compare its energy spectrum and ground-state behavior to the classical oscillator.

Introduction

Chapter 7 established that a wave packet, not a point trajectory, is the appropriate description of a quantum particle, and that this wave nature is directly responsible for the Heisenberg uncertainty principle. This chapter introduces the equation that governs how such a wave evolves: the **Schrödinger equation**, proposed by Erwin Schrödinger in 1926. It plays the role in quantum mechanics that Newton's second law plays in classical mechanics — given a system's wave function at one instant and the forces (via a potential energy function) acting on it, the Schrödinger equation determines the wave function at all later times. Solving it for a sequence of increasingly realistic potentials — a particle confined to a box, a particle encountering a step or barrier, a particle in a parabolic potential — reveals features with no classical counterpart: quantized energy levels, a nonzero minimum energy, partial reflection where classical physics predicts certain transmission, and the ability of a particle to be found where classical mechanics says it cannot.

8.1 The Wave Function and the Schrödinger Equation

The Wave Function and Its Interpretation

Quantum mechanics represents a particle's state by a complex-valued function of position and time, $\Psi(x,t)$ (in one dimension), called the **wave function**. Max Born proposed the interpretation now universally adopted: $|\Psi(x,t)|^2\,dx$ is the **probability** of finding the particle between x and $x+dx$ at time t, if a position measurement is performed. Because the particle must be found *somewhere*, an acceptable wave function must be **normalized**:

$$\int_{-\infty}^{\infty} |\Psi(x,t)|^2\,dx = 1. \tag{8.1}$$

For Ψ to yield a sensible probability density, it (and, where the potential is finite, its first derivative) must be single-valued, finite, and continuous; discontinuities or divergences in Ψ would correspond to ill-defined or infinite probability densities.

> Notation used throughout this chapter: capital $\Psi(x,t)$ denotes the full, time-dependent wave function, while lowercase $\psi(x)$ denotes only its spatial part, introduced below once time dependence is separated out. The two are related by $\Psi(x,t) = \psi(x)e^{-iEt/\hbar}$ for a stationary state, but $\psi(x)$ alone does not solve the full time-dependent equation.

The Born rule is easy to state and easy to under-read, because probability is also what a classical physicist reaches for when describing ignorance about a definite fact. The two are not the same, and Figure 8.1 is built around the difference. Its first screen puts a classical coin under a cover beside a quantum one: both give heads half the time, but the classical coin already *is* heads or tails while it is hidden, and the quantum one is in a superposition of the two until it is looked at. Its later screens make the distinction operational — single photons through a polarizer arriving with probability $\cos^2\theta$, and spin-$\frac{1}{2}$ particles through a chain of analyzers whose statistics no assignment of pre-existing values reproduces.

Note

Max Born's probabilistic reading of Ψ, introduced almost as an aside in a 1926 paper on scattering theory, was deeply controversial even among the theory's founders. Schrödinger himself hoped $|\Psi|^2$ would describe a smoothly spread-out electron charge density rather than a probability distribution over discrete measurement outcomes, and Einstein never fully accepted it either, famously objecting that "God does not play dice with the universe." The interpretation nonetheless survived every experimental test it was put to, and Born received the Nobel Prize in Physics for it only in 1954 — nearly three decades after the fact, an unusually long wait for a result so central to the theory.

Expectation Values and Operators

Because $|\Psi(x,t)|^2$ gives only a probability distribution, not a definite trajectory, a quantum "measurement" of position generally yields different results on identically prepared systems, with a statistical average — the **expectation value** — defined by

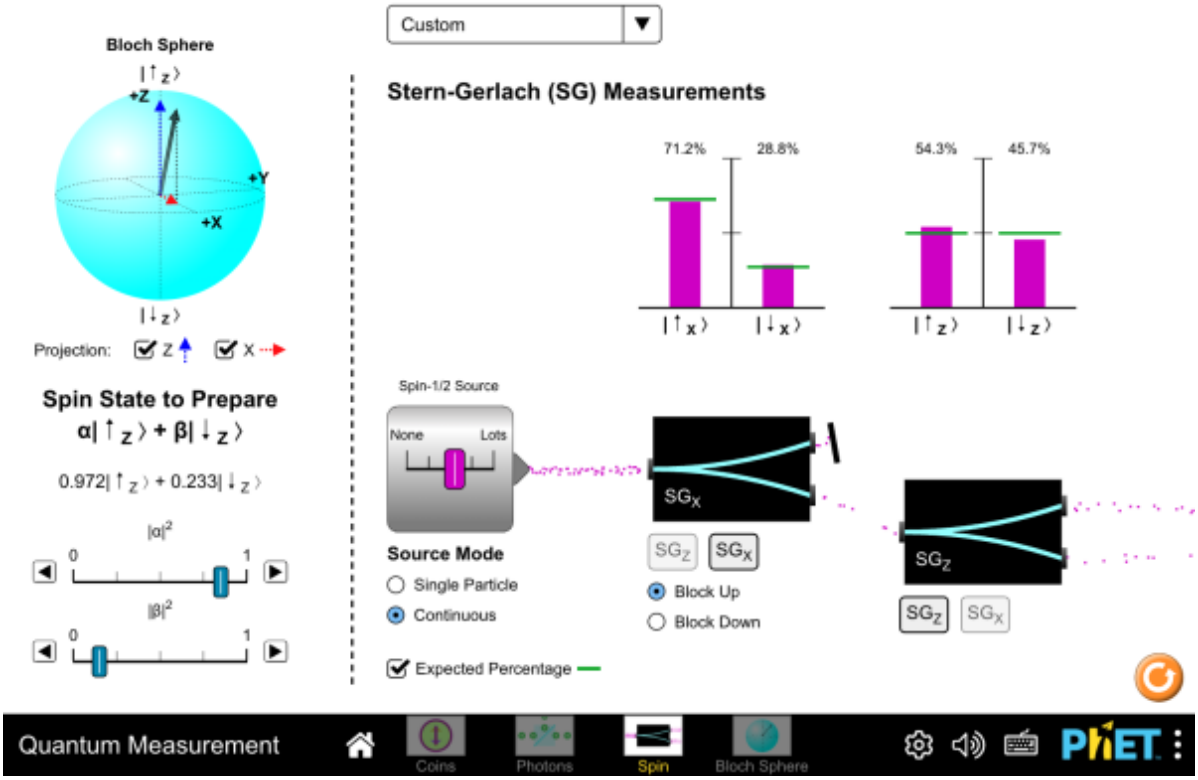

Figure 8.1: Measurement statistics for a classical coin, for single photons through a polarizer, and for spin-$\frac{1}{2}$ particles. $|\Psi|^2$ is a probability, but not a probability about something already decided — the distinction this chapter's mathematics quietly assumes. Interactive simulation: Quantum Measurement

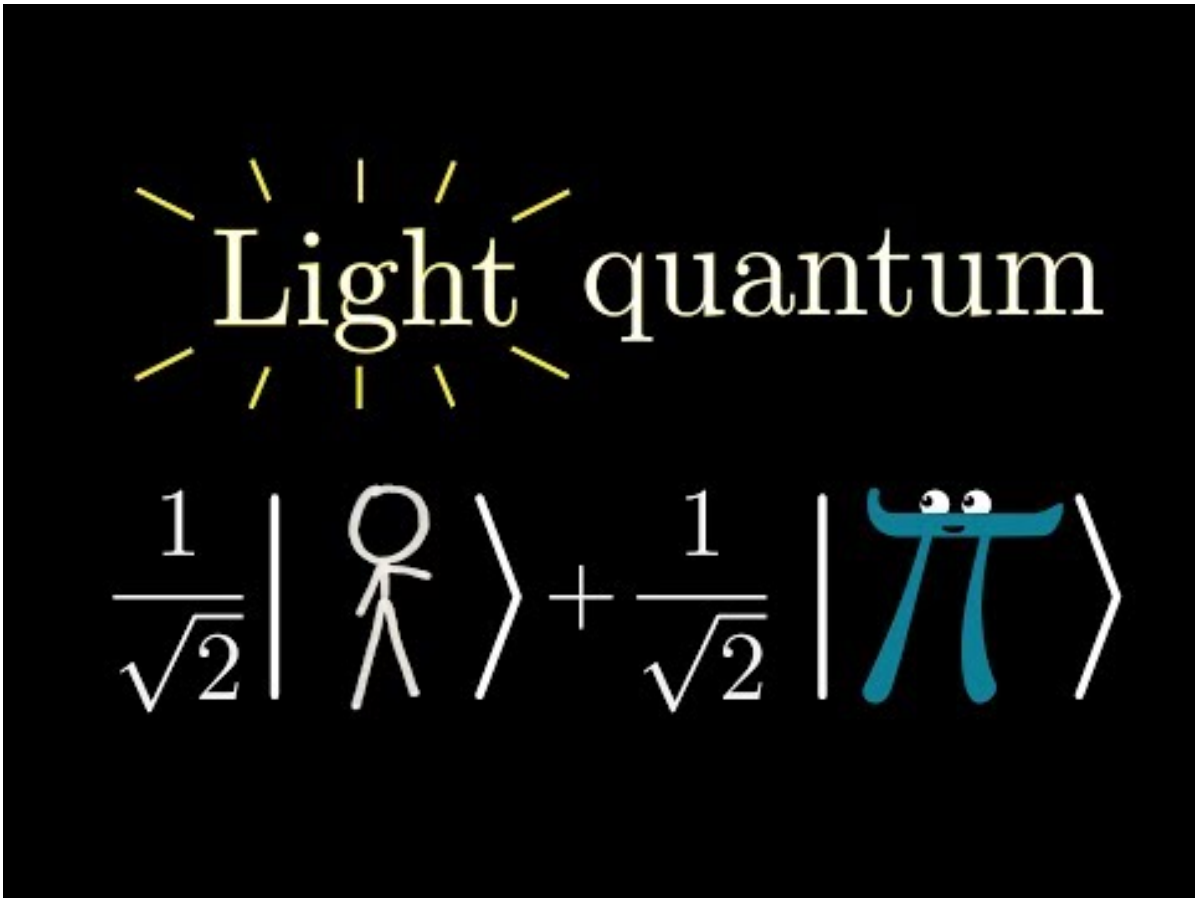

Figure 8.2: A first-principles, math-forward companion to the Born rule above, building the same superposition idea up from state vectors rather than from the coin-and-polarizer examples. Watch on YouTube: Some light quantum mechanics (with minutephysics)

$$\langle x \rangle = \int_{-\infty}^{\infty} \Psi^*(x,t)\, x\, \Psi(x,t)\, dx, \tag{8.2}$$

the same weighted-average construction used for any probability distribution in statistics, here weighted by $|\Psi|^2$. Momentum, energy, and other physical quantities are represented in quantum mechanics by **operators** acting on Ψ — most importantly the momentum operator $\hat{p} = -i\hbar\,\partial/\partial x$, already implicit in the correspondence used to motivate the Schrödinger equation below — and their expectation values follow the same sandwich pattern,

$$\langle p \rangle = \int_{-\infty}^{\infty} \Psi^*(x,t) \left(-i\hbar \frac{\partial}{\partial x}\right) \Psi(x,t)\, dx. \tag{8.3}$$

Expectation values are the quantities that connect the abstract wave function to numbers that can be compared with experiment: repeating a position (or momentum) measurement on many identically prepared copies of the same system and averaging the results reproduces $\langle x \rangle$ (or $\langle p \rangle$), while the statistical spread of those repeated measurements is exactly the Δx (or Δp) appearing in the Heisenberg uncertainty principle of Chapter 7.

> The operators representing physical observables — position x, momentum $-i\hbar\,\partial/\partial x$, and others introduced in later chapters — are all **Hermitian**. That property is exactly what guarantees expectation values like $\langle x \rangle$ and $\langle p \rangle$ always come out real, as a measurable quantity must, even though Ψ itself is complex.

Ehrenfest's Theorem: Recovering Newton's Second Law on Average

Although an individual quantum particle does not follow a definite classical trajectory, its *expectation values* obey equations strikingly close to the classical equations of motion. Differentiating $\langle x \rangle$ and $\langle p \rangle$ with respect to time and using the time-dependent Schrödinger equation to evaluate the results (a calculation carried out in more advanced treatments) gives **Ehrenfest's theorem**,

$$\frac{d\langle x \rangle}{dt} = \frac{\langle p \rangle}{m}, \qquad \frac{d\langle p \rangle}{dt} = -\left\langle \frac{dV}{dx} \right\rangle = \langle F \rangle, \tag{8.4}$$

which are exactly Newton's second law, $F = ma$, written for expectation values rather than for a sharp classical position and momentum. This is a precise, quantitative statement of the correspondence principle developed later in this chapter: quantum mechanics does not contradict Newtonian mechanics but contains it, recovered as a statement about the *average* behavior of a quantum ensemble. For a wave packet narrow enough that $\langle dV/dx \rangle \approx dV/dx|_{x=\langle x \rangle}$ (a good approximation whenever the potential varies slowly across the packet's width, typically satisfied for macroscopic objects but not necessarily for an electron confined to atomic dimensions), Ehrenfest's theorem shows that the packet's center moves, to excellent approximation, exactly like a classical particle obeying Newton's second law — which is precisely why macroscopic objects, whose wave packets are always narrow on any macroscopically relevant length scale, appear to move along sharp classical trajectories even though their underlying description is fully quantum mechanical.

Deriving $d\langle x \rangle/dt = \langle p \rangle/m$ from the Schrödinger equation

Differentiating $\langle x \rangle = \int \Psi^* x \Psi\, dx$ under the integral sign gives

$$\frac{d\langle x\rangle}{dt} = \int \frac{\partial \Psi^*}{\partial t}\, x\, \Psi\, dx + \int \Psi^*\, x\, \frac{\partial \Psi}{\partial t}\, dx. \tag{8.5}$$

Solving the time-dependent Schrödinger equation for $\partial\Psi/\partial t$, and its complex conjugate for $\partial\Psi^*/\partial t$ (using that V is real), gives

$$\frac{\partial \Psi}{\partial t} = \frac{i\hbar}{2m}\frac{\partial^2 \Psi}{\partial x^2} - \frac{i}{\hbar} V\Psi, \qquad \frac{\partial \Psi^*}{\partial t} = -\frac{i\hbar}{2m}\frac{\partial^2 \Psi^*}{\partial x^2} + \frac{i}{\hbar} V\Psi^*. \tag{8.6}$$

Substituting both into the expression for $d\langle x\rangle/dt$, the two terms containing V are identical in magnitude and opposite in sign ($\Psi^* V x \Psi$ either way, since V and x are ordinary multiplying functions), so they cancel exactly, leaving only

$$\frac{d\langle x\rangle}{dt} = \frac{i\hbar}{2m}\left[\int \Psi^*\, x\, \frac{\partial^2 \Psi}{\partial x^2}\, dx - \int \frac{\partial^2 \Psi^*}{\partial x^2}\, x\, \Psi\, dx\right]. \tag{8.7}$$

Integrating each term by parts twice in x — the boundary terms vanish because a normalizable Ψ (and its derivative) must go to zero as $x \to \pm\infty$ — collapses the bracket to $-2\int \Psi^*\,(\partial\Psi/\partial x)\, dx$, so that

$$\frac{d\langle x\rangle}{dt} = -\frac{i\hbar}{m}\int \Psi^* \frac{\partial \Psi}{\partial x}\, dx = \frac{1}{m}\int \Psi^* \left(-i\hbar \frac{\partial}{\partial x}\right) \Psi\, dx = \frac{\langle p\rangle}{m}, \tag{8.8}$$

recovering the first of Ehrenfest's equations directly from the postulated wave equation, with no additional assumption beyond normalizability. The second equation, $d\langle p\rangle/dt = \langle F\rangle$, follows from the same procedure applied to $\langle p\rangle$ instead of $\langle x\rangle$.

The Time-Dependent Schrödinger Equation

For a particle of mass m moving in one dimension under a potential energy $V(x,t)$, the wave function obeys the **time-dependent Schrödinger equation**:

$$i\hbar \frac{\partial \Psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2 \Psi(x,t)}{\partial x^2} + V(x,t)\,\Psi(x,t). \tag{8.9}$$

This equation is postulated, not derived from more elementary principles — its justification, as with Newton's laws, is that its predictions match experiment. It can, however, be motivated heuristically: substituting a free-particle plane wave $\Psi \propto e^{i(kx-\omega t)}$ (a wave of definite momentum $p = \hbar k$, per de Broglie, and definite energy $E = \hbar\omega$, per Planck–Einstein) and comparing to the classical nonrelativistic energy relation $E = p^2/2m + V$ reproduces exactly the operator correspondences $E \to i\hbar\,\partial/\partial t$ and $p \to -i\hbar\,\partial/\partial x$ built into the equation.

The Free Particle and the Need for Wave Packets

Setting $V = 0$, the plane wave $\Psi(x,t) = Ae^{i(kx-\omega t)}$ used above to motivate the equation is itself an exact solution, with $\omega = \hbar k^2/2m$ (the same nonrelativistic dispersion relation used in Chapter 7 to derive the group velocity). This solution has a serious defect as a description of an actual free particle, however: $|\Psi|^2 = |A|^2$ is the same at every point x and every time t, so it cannot be normalized ($\int |\Psi|^2 dx$ diverges) and represents a particle equally likely to be found anywhere in all of space — a definite momentum but no localization whatsoever, the extreme limit of the uncertainty principle in which $\Delta p = 0$ forces $\Delta x \to \infty$. A physically realizable free particle is instead represented, exactly as in Chapter 7, by

a normalizable **wave packet**: a superposition of plane-wave solutions of many different k, each individually a solution of the time-dependent Schrödinger equation (which is linear, so any superposition of solutions is itself a solution), combined so that the total wave function is localized. Because different k-components have different $\omega(k)$ and therefore different phase and group velocities, such a packet is precisely the object that spreads over time, as discussed in Chapter 7; the Schrödinger equation makes that spreading a specific, calculable prediction rather than a qualitative expectation.

Stationary States and the Time-Independent Equation

When the potential $V(x)$ does not depend on time, the Schrödinger equation admits **separable** solutions of the form $\Psi(x,t) = \psi(x)e^{-iEt/\hbar}$, where $\psi(x)$ satisfies the **time-independent Schrödinger equation**:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + V(x)\,\psi(x) = E\psi(x). \tag{8.10}$$

Such solutions are called **stationary states**: although $\Psi(x,t)$ itself oscillates in time through the phase factor $e^{-iEt/\hbar}$, the probability density $|\Psi(x,t)|^2 = |\psi(x)|^2$ is time-independent, and E is the definite, sharply-valued energy of the state. Because the time-independent equation is a linear, second-order differential equation with boundary conditions imposed by the requirement that ψ be normalizable, it typically admits solutions — and hence allowed values of E — only for a discrete set of energies when the particle is confined (bound) by the potential. This is the origin of **energy quantization** in quantum mechanics: not an assumption added by hand, as in the Bohr model, but a direct mathematical consequence of solving a boundary-value problem for a confined wave. When the particle is *not* confined (as in the potential step below, where the particle can escape to $x \to +\infty$), the same equation instead admits solutions for a continuous range of E, describing **scattering states** rather than bound states.

8.2 Bound and Scattering States in One Dimension

The Infinite Square Well

The simplest confining potential is the **infinite square well**: $V(x) = 0$ for $0 < x < L$, and $V(x) = \infty$ elsewhere, representing a particle strictly confined to a box of width L (an idealization of, e.g., an electron trapped between strong barriers). Since ψ must vanish wherever $V = \infty$ (an infinite potential permits zero probability of the particle being found there), the boundary conditions are $\psi(0) = \psi(L) = 0$.

> **Tip**
> Before matching boundary conditions at any interface, check whether the potential jump there is finite or infinite. Continuity of ψ' (used below for the potential step and the finite well) is derived by integrating the Schrödinger equation across the boundary, a step that requires V to stay finite there; it does not apply at an infinite wall. That is why the infinite well uses only $\psi(0) = \psi(L) = 0$ — ψ' is in fact discontinuous at each wall, jumping from a finite value inside to zero outside, where $\psi \equiv 0$ identically.

Inside the well, the time-independent equation reduces to $\psi'' = -(2mE/\hbar^2)\psi$, with general solution $\psi(x) = A\sin(kx) + B\cos(kx)$, $k \equiv \sqrt{2mE}/\hbar$. The condition $\psi(0) = 0$ forces $B = 0$; the condition $\psi(L) = 0$ then forces $\sin(kL) = 0$, i.e. $kL = n\pi$ for a positive integer $n = 1, 2, 3, \ldots$ ($n = 0$ is excluded, since it gives $\psi \equiv 0$ everywhere — no particle at all). Solving for the allowed energies,

$$E_n = \frac{n^2\pi^2\hbar^2}{2mL^2} = \frac{n^2h^2}{8mL^2}, \qquad n = 1, 2, 3, \ldots \tag{8.11}$$

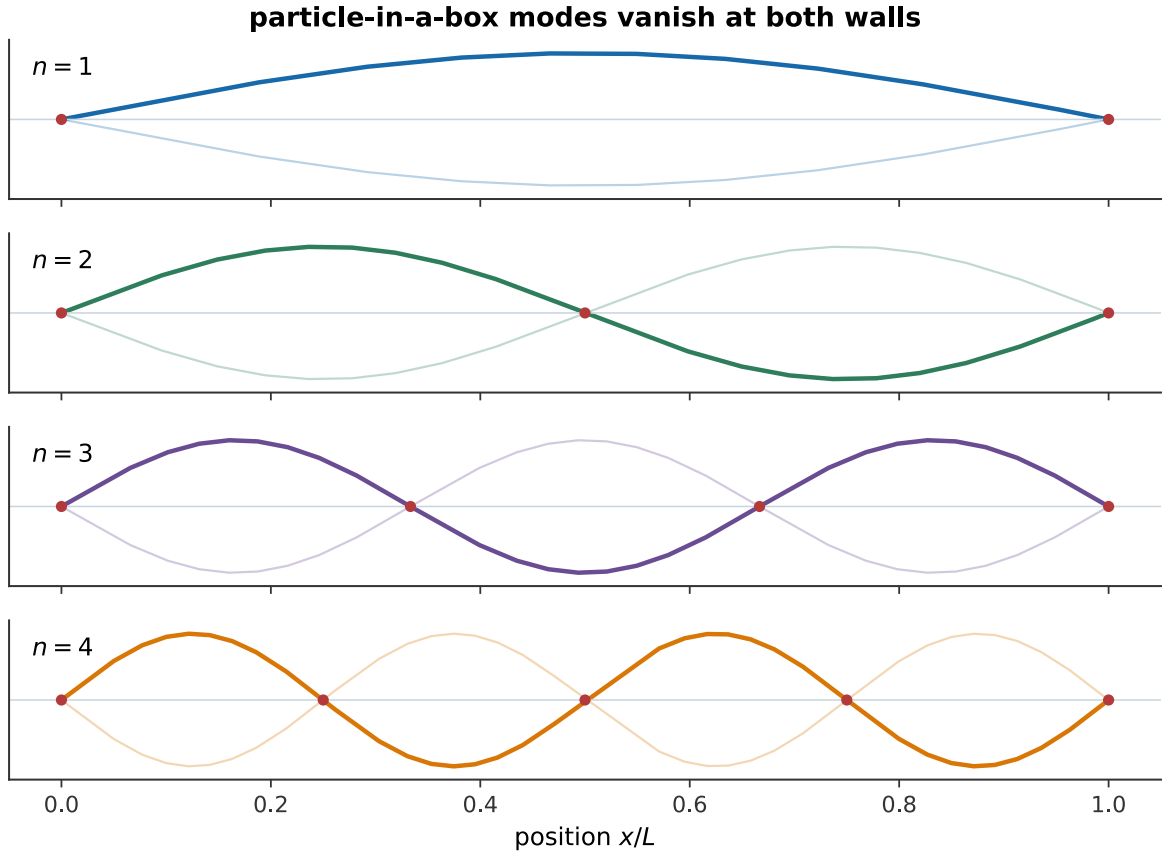

Figure 8.3: The first four modes of a wave confined between fixed boundaries. Read the transverse displacement as $\psi(x)$ rather than as a literal string: the boundary condition is identical, $k_n = n\pi/L$, and the increasing number of nodes is the particle-in-a-box quantum number made visible.

Normalizing $\psi_n(x) = A\sin(n\pi x/L)$ over $[0, L]$ gives $A = \sqrt{2/L}$. Several features have no classical analog:

- **Energy is quantized**, in discrete levels $E_n \propto n^2$, rather than continuous as for a classical particle bouncing in a box.
- **The ground state ($n = 1$) has** $E_1 = \pi^2\hbar^2/2mL^2 > 0$, not zero. A particle strictly confined to a box, unlike a classical particle, can never be perfectly at rest — a direct consequence of the uncertainty principle: confining $\Delta x \sim L$ forces $\Delta p \gtrsim \hbar/L$, hence a minimum kinetic energy $\sim \hbar^2/mL^2$, of the same order as E_1.
- **The probability density $|\psi_n(x)|^2$ has $n-1$ interior nodes** (points where the particle has zero probability of being found) — for $n > 1$, positions strictly inside the box that

the particle can never occupy, again with no classical counterpart.

All three features are on display in Figure 8.4, which solves the time-independent equation numerically for a potential you choose and draws the resulting eigenfunctions on top of the level diagram. Start with a square well and make it deep: the levels approach the n^2 ladder derived above, the ground state sits visibly above the bottom of the well, and ψ_n picks up one more node with each step up that ladder. The same simulation carries the finite well, the harmonic oscillator, and the one-dimensional Coulomb potential, so it is worth returning to as those appear later in this chapter and in Chapter 10; the differences between their level *spacings* — n^2 here, uniform for the oscillator, $-1/n^2$ for Coulomb — are the fingerprints of the three potentials.

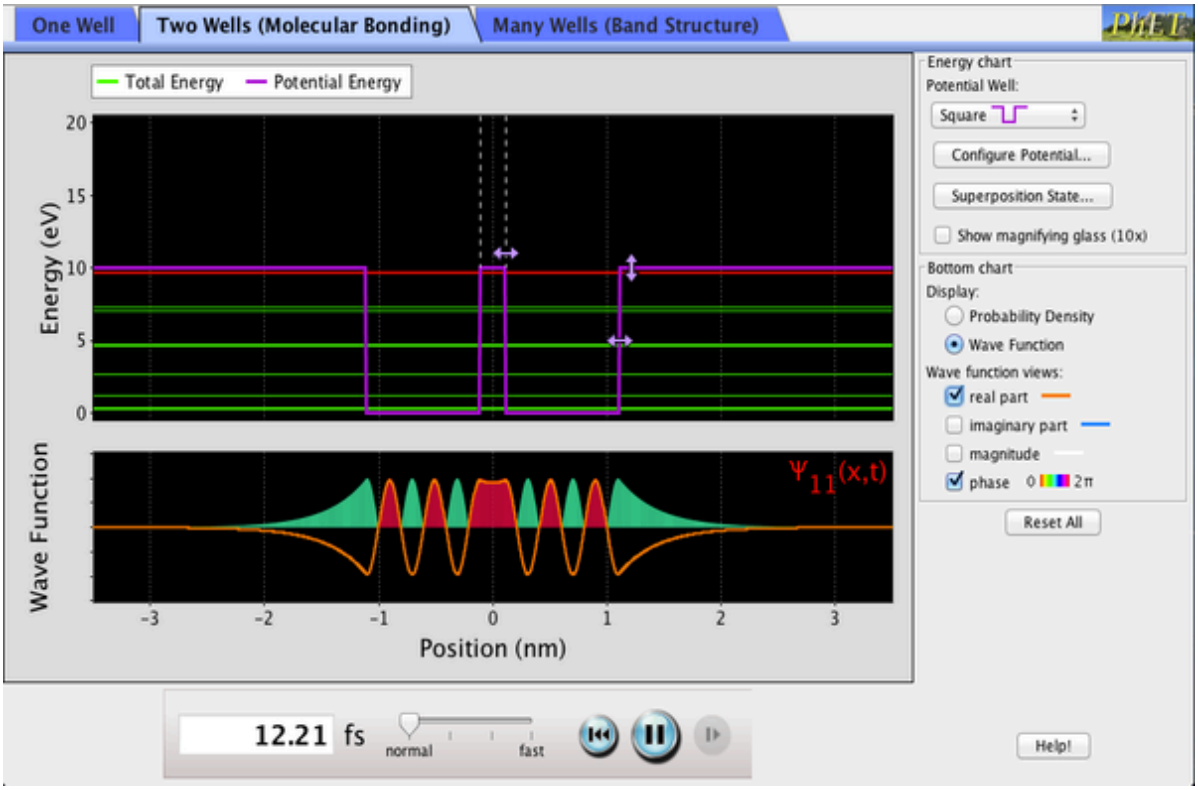

Figure 8.4: Bound states of a one-dimensional potential, with the well shape, depth, and width adjustable and the eigenfunctions drawn at their own energies. Superpositions of two eigenstates can be built and run in time, which is where the stationary states stop being stationary. Interactive simulation: Quantum Bound States

Worked Example: Expectation Values in the Ground State

For the infinite-well ground state $\psi_1(x) = \sqrt{2/L}\sin(\pi x/L)$, the probability density $|\psi_1(x)|^2$ is symmetric about the well's midpoint $x = L/2$, so by symmetry alone,

$$\langle x \rangle = \frac{L}{2}, \tag{8.12}$$

without needing to evaluate the defining integral explicitly — the particle is, on average, found at the center of the well, exactly as classical intuition would suggest, even though (unlike a classical particle) it is never actually found moving through the center with a definite

velocity. By the same symmetry, $\langle p \rangle = 0$: the stationary state carries no net probability current in either direction. However, $\langle p^2 \rangle \neq 0$: since the ground-state energy is purely kinetic ($V = 0$ inside the well), $E_1 = \langle p^2 \rangle / 2m$, so $\langle p^2 \rangle = 2mE_1 = \pi^2\hbar^2/L^2 \neq 0$. The particle has a nonzero *typical* momentum magnitude, $\sqrt{\langle p^2 \rangle} = \pi\hbar/L$, even though its *average* momentum is zero — a direct, quantitative illustration of the uncertainty principle at work: a particle confined to width $\Delta x \sim L$ cannot have $p = 0$ exactly, only $\langle p \rangle = 0$ with a nonzero spread $\Delta p = \sqrt{\langle p^2 \rangle - \langle p \rangle^2} = \pi\hbar/L$, consistent with $\Delta x \, \Delta p \sim \hbar$.

The Potential Step: Partial Reflection and Transmission

Before turning to bound states in a finite well, it is instructive to consider a simpler, unbound configuration: a **potential step**, $V(x) = 0$ for $x < 0$ and $V(x) = V_0$ for $x > 0$, with a particle of energy $E > V_0$ incident from $x = -\infty$. Classically, the particle simply slows down (its kinetic energy dropping from E to $E - V_0$) as it crosses $x = 0$, but continues forward with certainty — transmission probability 1, reflection probability 0.

Quantum mechanically, the time-independent equation in each region gives oscillatory solutions, $\psi_{\mathrm{I}}(x) = Ae^{ik_1 x} + Be^{-ik_1 x}$ for $x < 0$ (an incident wave of amplitude A plus a reflected wave of amplitude B) and $\psi_{\mathrm{II}}(x) = Ce^{ik_2 x}$ for $x > 0$ (a transmitted wave only, since nothing is incident from $x = +\infty$), with $k_1 = \sqrt{2mE}/\hbar$ and $k_2 = \sqrt{2m(E - V_0)}/\hbar$. Matching ψ and ψ' continuously at $x = 0$ gives two equations in A, B, C, which solve to

$$\frac{B}{A} = \frac{k_1 - k_2}{k_1 + k_2}, \qquad \frac{C}{A} = \frac{2k_1}{k_1 + k_2}. \tag{8.13}$$

Because $B \neq 0$ whenever $k_1 \neq k_2$ (i.e., whenever $V_0 \neq 0$), **some of the incident wave is reflected even though** $E > V_0$ — a purely wave-mechanical effect (the same kind of partial reflection that occurs when light passes from one medium of different refractive index to another, Chapter 4) with no classical particle counterpart at all. Comparing probability flux (rather than amplitude) on the two sides gives the **reflection and transmission coefficients**,

$$R = \left(\frac{k_1 - k_2}{k_1 + k_2} \right)^2, \qquad T = \frac{4k_1 k_2}{(k_1 + k_2)^2}, \qquad R + T = 1, \tag{8.14}$$

the last equality expressing overall conservation of probability: every incident particle is eventually found either reflected or transmitted. If instead $E < V_0$, then k_2 becomes imaginary, the "transmitted" solution turns into a decaying exponential (no propagating wave in region II at all), and $R = 1$ exactly — total reflection, as classically expected — but, just as in the wave-packet argument of Chapter 7, the wave function does not vanish abruptly at $x = 0$; writing $k_2 = i\kappa$ with $\kappa = \sqrt{2m(V_0 - E)}/\hbar$ real, the wave function in region II becomes $\psi_{\mathrm{II}}(x) \propto e^{-\kappa x}$, decaying with a characteristic **penetration depth** $1/\kappa$ rather than vanishing abruptly at $x = 0$. The particle therefore has a small but nonzero probability of being found some distance *into* the classically forbidden region, even though it is certain, eventually, to be reflected back the way it came — a preview of the tunneling phenomenon developed below, in which a *second* boundary, ending the forbidden region before ψ has fully decayed away, allows the particle to escape entirely rather than merely penetrate and return.

Both halves of that story — partial reflection at a step the particle has the energy to cross, and exponential decay into a step it does not — can be run against a real wave packet rather than a plane wave in Figure 8.5. Send a packet at a step with $E > V_0$ and it visibly splits: part of it continues, slower and stretched, and part comes back, with the areas under the two pieces reproducing the T and R computed above. Lower the energy below V_0 and the packet is

entirely reflected, but during the encounter it leaks a decaying tail into the barrier. Replacing the step by a barrier of finite thickness — the simulation's next potential — is the whole of tunneling: the tail reaches the far side before it has died away, and what emerges there is a transmitted packet.

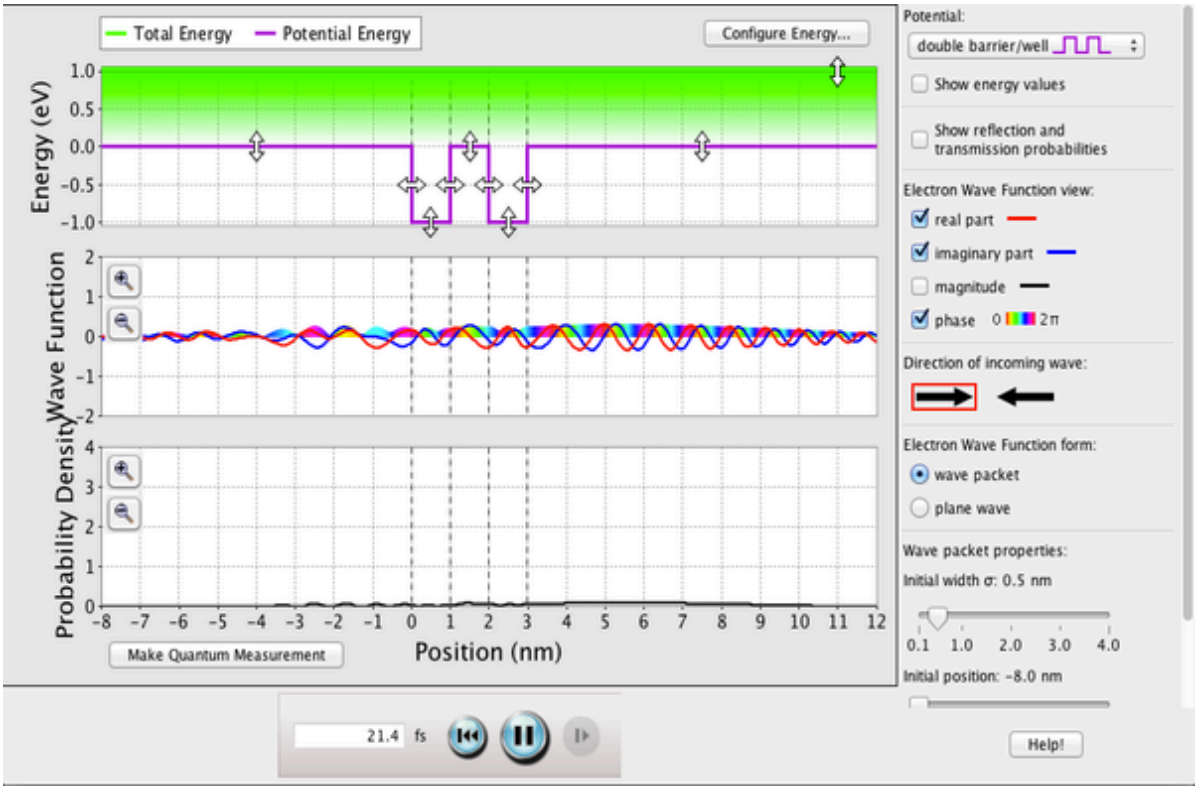

Figure 8.5: A wave packet incident on a step, a barrier, or a double barrier, with the real and imaginary parts of ψ and the probability density all available. The energy relative to the barrier height is adjustable, as are the barrier's width and height — the two parameters the transmission coefficient of the next section depends on exponentially. Interactive simulation: Quantum Tunneling and Wave Packets

Worked Example: Reflection at a Potential Step

An electron with kinetic energy $E = 2.00$ eV approaches a potential step of height $V_0 = 1.00$ eV. Since $k \propto \sqrt{E}$ (with the same proportionality constant on both sides, as m is unchanged), the ratio $k_2/k_1 = \sqrt{(E - V_0)/E} = \sqrt{(1.00\text{ eV})/(2.00\text{ eV})} = 0.707$. The reflection coefficient is then

$$R = \left(\frac{1 - 0.707}{1 + 0.707}\right)^2 = \left(\frac{0.293}{1.707}\right)^2 = 0.0295, \tag{8.15}$$

so about 3% of an incident beam of such electrons is reflected by the step, and $T = 1 - R \approx 0.97$ is transmitted — a small but entirely real and measurable effect, with no analog for a classical particle rolling over an upward step in a potential energy landscape.

8.3 Tunneling, the Oscillator, and Correspondence

The Finite Square Well and Quantum Tunneling

A more physically realistic bound-state model replaces the infinitely high walls of the square well with walls of finite height V_0: $V(x) = 0$ for $0 < x < L$ and $V(x) = V_0$ outside. For a bound state with $E < V_0$, the time-independent equation outside the well becomes $\psi'' = +\kappa^2\psi$ with $\kappa \equiv \sqrt{2m(V_0 - E)}/\hbar$ real, whose normalizable solutions are decaying exponentials, $\psi(x) \propto e^{-\kappa|x|}$ moving away from the well, rather than the oscillatory sines and cosines found inside.

> For order-of-magnitude estimates in atomic and nuclear problems, it is often quicker to carry $\hbar c \approx 197.3$ eV · nm = 197.3 MeV · fm as a single combined constant than to track $\hbar$ and c separately — it turns a decay constant like $\kappa = \sqrt{2m(V_0 - E)}/\hbar$ into an expression in rest energies mc^2 and $(V_0 - E)$ alone, the trick used in several worked examples and problems below.

This is the central qualitative difference from the infinite well: **the wave function does not vanish at the walls, but decays exponentially into the classically forbidden region** where $E < V(x)$ — a region a classical particle could never enter, since it would require negative kinetic energy there. Quantum mechanically, there is a small but nonzero probability of finding the particle just outside the well. Matching ψ and ψ' continuously at each wall (rather than forcing $\psi = 0$ as in the infinite well) yields a transcendental equation for the allowed energies, which must generally be solved numerically or graphically. Choosing the well symmetrically, $V(x) = 0$ for $-L/2 < x < L/2$, exploits the potential's reflection symmetry to separate the bound-state solutions into two families: **even** solutions, $\psi(x) = A\cos(kx)$ inside the well ($k \equiv \sqrt{2mE}/\hbar$), for which matching to the exterior decay $\psi \propto e^{-\kappa|x|}$ gives the condition

$$k\tan(kL/2) = \kappa, \tag{8.16}$$

and **odd** solutions, $\psi(x) = A\sin(kx)$ inside the well, for which the matching condition is instead

$$-k\cot(kL/2) = \kappa. \tag{8.17}$$

Because $\kappa = \sqrt{2m(V_0 - E)}/\hbar$ and k are both functions of the single unknown E, each of these is a single transcendental equation in E, most easily solved graphically by plotting both sides as functions of k and reading off the intersections; each intersection is one allowed bound-state energy. The lowest bound state is always of the even type (a symmetric, node-free wave function, exactly like the infinite well's ground state); qualitatively, the finite well has *fewer* bound states than the infinite well of the same width, although a one-dimensional attractive finite well always has at least one bound state for any nonzero depth. Each allowed energy is slightly lower than the corresponding infinite-well value, because the wave function's penetration into the forbidden region effectively widens the well.

This same exponential penetration underlies **quantum tunneling**: if a particle of energy E encounters a finite-width barrier of height $V_0 > E$ (rather than an infinite wall), the wave function decays but does not necessarily reach zero before the barrier ends, re-emerging on the far side as a (reduced-amplitude, but nonzero) oscillatory wave. There is then a nonzero probability — the **transmission coefficient** T — that the particle is found on the far side of

the barrier, despite lacking, classically, enough energy to pass over it. For a barrier of width L and height $V_0 > E$, in the regime where the barrier strongly suppresses transmission ($\kappa L \gg 1$),

$$T \approx e^{-2\kappa L}, \qquad \kappa = \frac{\sqrt{2m(V_0 - E)}}{\hbar}, \tag{8.18}$$

showing that tunneling probability falls off exponentially with both the barrier's width and the square root of $m(V_0 - E)$ — which is why tunneling is significant for light particles (electrons) through thin barriers but utterly negligible for macroscopic objects. Tunneling is not a mathematical curiosity; it is the mechanism behind alpha decay (Chapter 13), the scanning tunneling microscope, and (approximately) the operation of tunnel diodes.

Worked Example: Distance Sensitivity of a Scanning Tunneling Microscope

A scanning tunneling microscope (STM) images a conducting surface by scanning a sharp metal tip a fraction of a nanometer above it and measuring the tunneling current of electrons crossing the vacuum gap — a gap that acts as a potential barrier of height roughly equal to the metal's work function, here taken as $V_0 - E \approx 4.0$ eV. The decay constant is

$$\kappa = \frac{\sqrt{2m_e(V_0 - E)}}{\hbar} = \frac{\sqrt{2(9.11 \times 10^{-31}\ \text{kg})[(4.0\ \text{eV})(1.60 \times 10^{-19}\ \text{J/eV})]}}{1.055 \times 10^{-34}\ \text{J}\cdot\text{s}} \approx 1.02{\times}10^{10}\ \text{m}^{-1} \approx 10.2\ \text{nm}^{-1}. \tag{8.19}$$

Since $T \propto e^{-2\kappa L}$, increasing the tip–surface gap by just $\Delta L = 0.10$ nm (one angstrom) changes the transmission probability, and hence the measured tunneling current, by a factor of

$$e^{-2\kappa\,\Delta L} = e^{-2(10.2\ \text{nm}^{-1})(0.10\ \text{nm})} = e^{-2.05} \approx \frac{1}{7.8}, \tag{8.20}$$

roughly an order of magnitude per angstrom of vertical displacement — exactly the extreme sensitivity that allows an STM to resolve individual atoms on a surface by tracking tiny changes in tunneling current as the tip scans across atomic-scale height variations.

Worked Example: Order-of-Magnitude Estimate for Alpha Decay

Alpha decay (Chapter 13) is understood as an alpha particle ($m_\alpha \approx 6.64 \times 10^{-27}$ kg) tunneling through the Coulomb barrier confining it inside a heavy nucleus. As a rough, illustrative model, approximate this barrier as rectangular, with the alpha particle's energy falling $\Delta E \approx V_0 - E = 20$ MeV below the barrier height and an effective barrier width $L \approx 7$ fm $= 7{\times}10^{-15}$ m (real barriers are Coulombic, not rectangular, and a proper treatment integrates $\kappa(r)$ over the barrier's actual shape — the **Gamow factor** — but the rectangular approximation captures the essential physics). Then

$$\kappa = \frac{\sqrt{2m_\alpha \Delta E}}{\hbar} = \frac{\sqrt{2(6.64 \times 10^{-27}\ \text{kg})[(20\ \text{MeV})(1.602 \times 10^{-13}\ \text{J/MeV})]}}{1.055 \times 10^{-34}\ \text{J}\cdot\text{s}} \approx 1.96 \times 10^{15}\ \text{m}^{-1}, \tag{8.21}$$

so the tunneling probability per attempt is

$$T \approx e^{-2\kappa L} = e^{-2(1.96\times10^{15}\ \text{m}^{-1})(7\times10^{-15}\ \text{m})} = e^{-27.4} \approx 1 \times 10^{-12}. \tag{8.22}$$

This tiny number is not the final answer for a decay rate; it must be combined with an **assault frequency**, the rate at which the alpha particle, rattling back and forth inside the nucleus at a speed of order 10^7 m/s across a nuclear diameter of order 10^{-14} m, "attempts" to escape — roughly $f \sim 10^{21}$ s^{-1}. The decay rate is then of order $fT \sim (10^{21}\ \text{s}^{-1})(10^{-12}) = 10^{9}\ \text{s}^{-1}$, corresponding to a half-life of order 10^{-9} s for *this* choice of illustrative barrier parameters. Real alpha emitters span half-lives from microseconds to billions of years, entirely because T depends *exponentially* on ΔE and L: changing the barrier parameters by a modest amount changes T, and hence the half-life, by many orders of magnitude — exactly the empirical pattern captured by the **Geiger–Nuttall relation** revisited quantitatively in Chapter 13.

The Quantum Harmonic Oscillator

A particle in a potential $V(x) = \frac{1}{2}kx^2$ (with k here the spring constant, not a wave number) is the quantum analog of the classical simple harmonic oscillator, and it is important beyond this specific system because *any* smooth potential, expanded in a Taylor series about a point of stable equilibrium, is approximately parabolic near that minimum — the harmonic oscillator is the generic first approximation for small oscillations about equilibrium in essentially any bound system, including the vibrations of a diatomic molecule (Chapter 12).

> Throughout this section ω is the **angular** frequency (rad/s) of the corresponding classical oscillator, $\omega = \sqrt{k/m}$, while the *ordinary* frequency used for photon energies elsewhere in this book is $f = \omega/2\pi$. The worked examples below convert between the two, since a spring constant is naturally expressed via ω but spectroscopic data is usually quoted as f.

Solving the time-independent Schrödinger equation with this potential (the details require either a power-series method or an elegant operator technique, developed in more advanced treatments) yields an evenly spaced energy spectrum,

$$E_n = \left(n + \tfrac{1}{2}\right)\hbar\omega, \qquad n = 0, 1, 2, \ldots, \qquad \omega \equiv \sqrt{k/m}, \tag{8.23}$$

with ω the classical angular frequency of the corresponding classical oscillator. Two features stand out. First, unlike the square well, the spacing between adjacent levels, $\hbar\omega$, is the *same* for every n — a distinctive signature of the parabolic potential. Second, the ground state ($n = 0$) has energy $E_0 = \frac{1}{2}\hbar\omega \neq 0$, called the **zero-point energy**: even in its lowest possible energy state, a quantum oscillator retains irreducible energy and motion, again a manifestation of the uncertainty principle (a particle at rest, at the exact bottom of the well, would have $\Delta x = \Delta p = 0$, forbidden by $\Delta x\,\Delta p \geq \hbar/2$).

> **Warning**
> It is tempting to picture the zero-point energy $E_0 = \frac{1}{2}\hbar\omega$ as ordinary thermal-like jiggling that persists even at absolute zero, and then to wonder whether it could be tapped as a source of free energy or whether it violates the third law of thermodynamics. Neither concern applies: E_0 is the *irreducible* energy of a single stationary state, present even though $\langle p\rangle = 0$ in that state (exactly as for the infinite well's ground state) — there is no net directed motion to extract as work, and no colder state available for heat to flow toward. Zero-point energy is a firm floor set by $\Delta x\,\Delta p \geq \hbar/2$, not leftover thermal agitation that a sufficiently cold environment could remove.

The ground-state wave function, $\psi_0(x) \propto e^{-m\omega x^2/2\hbar}$, is a Gaussian, peaked (unlike the classical oscillator, which spends most of its time near the turning points, where it moves slowest) at the center $x = 0$ — another qualitative divergence from classical intuition that only disappears, via the correspondence principle, for large n, where the quantum probability distribution begins to average out to resemble the classical one.

Excited harmonic-oscillator wave functions follow the same node-counting pattern already seen for the infinite well: $\psi_n(x)$ has exactly n nodes (points, other than the boundaries at infinity, where the wave function crosses zero), so ψ_0 (the Gaussian ground state) has none, ψ_1 has one node at $x = 0$, ψ_2 has two nodes symmetric about the origin, and so on — a general feature of one-dimensional bound-state wave functions (also true of the infinite well, where ψ_n has $n - 1$ *interior* nodes in addition to the two required zeros at the walls themselves) that provides a quick, qualitative check on whether a proposed solution corresponds to the ground state, first excited state, or a higher state, without needing to solve the full equation.

Transitions between harmonic-oscillator levels by absorption or emission of a photon obey the **selection rule** $\Delta n = \pm 1$: an oscillating dipole (the physical mechanism by which a vibrating charged system couples to the electromagnetic field) can only connect states whose quantum numbers differ by exactly one, so the emitted or absorbed photon energy is always $\hbar\omega$, the same fixed spacing regardless of which pair of adjacent levels is involved — a fact used directly in Chapter 12 to interpret the vibrational spectra of diatomic molecules, where (to the extent the harmonic approximation holds) all vibrational transitions cluster near a single characteristic frequency rather than spreading across many distinct energies.

Worked Example: Zero-Point Motion, Macroscopic Versus Molecular

The characteristic spread of the ground-state Gaussian wave function is $\Delta x = \sqrt{\hbar/2m\omega}$. For a macroscopic oscillator — a 1.0 kg mass on a spring with $k = 100$ N/m, so $\omega = \sqrt{k/m} = 10$ rad/s — this gives

$$\Delta x = \sqrt{\frac{1.055 \times 10^{-34}\ \text{J}\cdot\text{s}}{2(1.0\ \text{kg})(10\ \text{s}^{-1})}} \approx 2.3 \times 10^{-18}\ \text{m}, \tag{8.24}$$

many orders of magnitude smaller than even a proton's radius ($\sim 10^{-15}$ m): zero-point motion is utterly unobservable for a macroscopic object, consistent with everyday experience that a spring at rest simply looks at rest. By contrast, a carbon monoxide molecule vibrating with $\omega \approx 4.0 \times 10^{14}$ rad/s (a typical molecular vibrational frequency, corresponding to the reduced mass $\mu \approx 1.14 \times 10^{-26}$ kg and bond stiffness $k \approx 1860$ N/m of the C–O bond) has

$$\Delta x = \sqrt{\frac{1.055 \times 10^{-34}\ \text{J}\cdot\text{s}}{2(1.14 \times 10^{-26}\ \text{kg})(4.0 \times 10^{14}\ \text{s}^{-1})}} \approx 3.4 \times 10^{-12}\ \text{m} = 3.4\ \text{pm}, \tag{8.25}$$

a few percent of the bond's equilibrium length (≈ 113 pm) — small, but not at all negligible, and precisely why zero-point vibrational motion has measurable consequences (a nonzero zero-point vibrational energy that must be included in molecular bond-energy calculations) for real molecules even though the analogous effect is entirely unobservable for a spring on a lab bench.

The Correspondence Principle Across Three Systems

The three systems solved in this chapter — the infinite well, the finite well with tunneling, and the harmonic oscillator — share a common thread despite their different mathematical

detail: each has a nonzero, quantized ground-state energy set by the uncertainty principle, the finite well and oscillator allow the particle to be found in regions forbidden to a classical particle of the same energy, and each converges toward classical predictions in an appropriate limit. For the infinite well, closely spaced levels at large n (where $E_n \propto n^2$ but the *fractional* spacing $\Delta E_n/E_n \to 0$) approach a quasi-continuous classical energy spectrum; for the harmonic oscillator, the large-n probability distribution $|\psi_n(x)|^2$ develops peaks near the classical turning points, where a classical oscillator spends most of its time, rather than the single central peak seen for $n = 0$. This is Bohr's **correspondence principle**: quantum predictions must merge smoothly into classical ones in the limit of large quantum numbers (equivalently, macroscopic action large compared to $\hbar$), and it serves as a valuable consistency check on any quantum-mechanical solution.

The same principle explains why quantum effects are unobservable for everyday macroscopic objects even though, formally, every object obeys the Schrödinger equation. A macroscopic pendulum or block on a spring has an enormous effective quantum number n (its total energy, measured in units of $\hbar\omega$, is astronomically large for ordinary masses, spring constants, and amplitudes), placing it deep in the correspondence-principle regime where quantized energy levels are spaced far too closely, relative to the total energy, to be distinguished from a continuum, and where the wave-packet spreading of Chapter 7 is negligible on any observable timescale. The three exactly solvable systems of this chapter are pedagogically valuable precisely because they are simple enough to solve exactly while still exhibiting, at small n, the full richness of specifically quantum behavior — behavior that becomes progressively harder to detect, though never strictly absent, as a system's size, mass, or energy grows toward the macroscopic scale.

See Also

Quantum Mechanics in Three Dimensions extends exactly this program — solve the time-independent Schrödinger equation subject to boundary conditions, and let quantization emerge from requiring a normalizable solution — to three dimensions, where the same separation-of-variables strategy used here for stationary states produces the quantum numbers that organize the hydrogen atom of Chapter 10.

8.4 Summary

- The wave function $\Psi(x,t)$ evolves according to the **time-dependent Schrödinger equation**; $|\Psi(x,t)|^2$ gives the probability density for finding the particle, per the Born interpretation, and Ψ must be normalized, single-valued, finite, and continuous. **Expectation values** of position and momentum are computed by sandwiching the position operator (x) or momentum operator $(-i\hbar\,\partial/\partial x)$ between Ψ^* and Ψ and integrating; **Ehrenfest's theorem** shows these expectation values obey Newton's second law, recovering classical mechanics as the average behavior of a narrow wave packet.
- For time-independent potentials, separable **stationary-state** solutions $\Psi = \psi(x)e^{-iEt/\hbar}$ satisfy the **time-independent Schrödinger equation**; requiring ψ to be normalizable in a confining potential generically forces E to take only discrete, quantized values, while an unconfined particle has a continuous spectrum of scattering states.
- The **infinite square well** gives $E_n = n^2h^2/8mL^2$, a nonzero ground-state energy, and wave functions with $n-1$ interior nodes — all with no classical counterpart.

- A **potential step** produces partial reflection even when the particle's energy exceeds the step height, with $R = \left(\frac{k_1 - k_2}{k_1 + k_2}\right)^2$ and $T = 1 - R$ — a purely wave-mechanical effect.
- The **finite square well** allows the wave function to penetrate into the classically forbidden region outside the well; the same mechanism, applied to a finite-width barrier, produces **quantum tunneling**, with transmission probability falling off exponentially with barrier width and $\sqrt{m(V_0 - E)}$, the basis of the scanning tunneling microscope's extreme distance sensitivity.
- The **quantum harmonic oscillator** has evenly spaced levels $E_n = (n + \frac{1}{2})\hbar\omega$, a selection rule $\Delta n = \pm 1$ for radiative transitions, and a nonzero **zero-point energy** $E_0 = \frac{1}{2}\hbar\omega$; it is the generic small-oscillation approximation to any smooth potential near a stable minimum.
- Bohr's **correspondence principle** — quantum predictions merge into classical ones at large quantum numbers — is a recurring consistency check across all three model systems.

8.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens on the standing-wave modes of an infinite well.

1. **Multiple choice, from a figure.** Why is the infinite-well ground-state energy not zero? Figure description: The first few standing-wave modes in a one-dimensional infinite square well. Choices: (A) A zero-momentum state cannot satisfy confinement and the well's boundary conditions. (B) The potential energy inside the well is infinite. (C) The particle must move at the speed of light. (D) The wave function is required to be constant everywhere.

2. **True or false.** A stationary state has a time-independent probability density even though its wave function carries a time-dependent phase.

3. **Drag the words.** Complete the interpretation of the wave function. The quantity $|\Psi|^2$ is the _____. A physical wave function must be _____. Bound-state energies arise from _____. Tunneling occurs because Ψ has _____. (Terms: probability density; normalizable; boundary conditions; nonzero amplitude beyond a finite barrier; probability amplitude squared only after measurement; zero amplitude inside every barrier.)

4. **Fill in the blanks.** Complete the infinite-well scaling. For a one-dimensional infinite well, En is proportional to n raised to the power _____, so E2/E1 = _____.

5. **Mark the words.** Mark the two statements that make tunneling possible. "A finite barrier permits an evanescent wave inside the classically forbidden region and the wave function can retain nonzero amplitude on the far side. The particle does not need energy greater than the barrier."

8.6 Conceptual Questions

1. Why is a wave function allowed to be complex even though every measured probability is real?

2. A stationary state's wave function has time dependence. Why is the state nevertheless called stationary?

3. Explain how boundary conditions, rather than a separate quantization rule, produce the discrete energies of a particle in a box.

4. Why can a quantum particle reflect from a potential step even when its energy exceeds the step height?

5. Why does increasing either the width or the height of a barrier suppress tunneling exponentially?

6. Compare the level spacing of the infinite well and harmonic oscillator. What does each reveal about its potential?

8.7 Problems

Exercise 8.1
(Straightforward)

An electron is confined to an infinite square well of width $L = 0.20$ nm (roughly an atomic diameter). Find (a) the ground-state energy E_1 in eV, and (b) the energy of the photon emitted in a transition from $n = 2$ to $n = 1$.

Exercise 8.2
(Straightforward)

Show, by direct substitution into the time-independent Schrödinger equation, that $\psi_n(x) = \sqrt{2/L}\sin(n\pi x/L)$ with $E_n = n^2\pi^2\hbar^2/2mL^2$ is indeed a solution for the infinite square well on $0 < x < L$.

Exercise 8.3
(Straightforward)

For the infinite-well state $\psi_2(x) = \sqrt{2/L}\sin(2\pi x/L)$ $(n = 2)$, use the symmetry of $|\psi_2(x)|^2$ about $x = L/2$ to state $\langle x\rangle$ without direct integration, and identify the location of the single interior node.

Exercise 8.4
(Straightforward)

An electron with kinetic energy $E = 3.00$ eV encounters a potential step of height $V_0 = 2.00$ eV. (a) Compute k_2/k_1. (b) Compute the reflection coefficient R and the transmission coefficient T. (c) Repeat for $V_0 = 0.500$ eV, and comment on how R changes as the step height decreases toward zero.

Exercise 8.5
(Straightforward)

Show algebraically, starting from the amplitude ratios $B/A = (k_1 - k_2)/(k_1 + k_2)$ and using $R = |B/A|^2$, $T = (k_2/k_1)|C/A|^2$ with $C/A = 2k_1/(k_1 + k_2)$, that $R + T = 1$ for the potential step.

Exercise 8.6

(Straightforward)

A proton with 5.0 MeV of kinetic energy strikes a rectangular potential barrier of height 10.0 MeV and width 2.0×10^{-15} m (roughly a nuclear dimension). (a) Compute $\kappa = \sqrt{2m(V_0 - E)}/\hbar$ for the proton in the barrier. (b) Estimate the tunneling transmission probability $T \approx e^{-2\kappa L}$. (c) Repeat for an alpha particle (mass four times the proton mass) under the same conditions and compare, explaining qualitatively why the heavier particle tunnels less readily.

Exercise 8.7

(Straightforward)

Using the STM worked example's value of $\kappa \approx 10.2$ nm^{-1}, find the factor by which the tunneling current changes if the tip–surface gap increases by 0.20 nm instead of 0.10 nm, and comment on why STM height measurements are typically precise to a small fraction of an angstrom.

Exercise 8.8

(Straightforward)

Verify that the ground-state wave function of the harmonic oscillator, $\psi_0(x) = A\,e^{-m\omega x^2/2\hbar}$, satisfies the time-independent Schrödinger equation for $V(x) = \tfrac{1}{2}m\omega^2 x^2$ with energy $E_0 = \tfrac{1}{2}\hbar\omega$, by direct substitution (you need not determine the normalization constant A).

Exercise 8.9

(Moderate)

A diatomic molecule vibrates approximately as a harmonic oscillator with classical frequency $f = \omega/2\pi = 8.7 \times 10^{13}$ Hz. Find (a) the zero-point energy in eV, and (b) the energy of a photon emitted in a transition between adjacent vibrational levels ($\Delta n = 1$), and (c) identify the region of the electromagnetic spectrum (see Chapter 6) in which this photon lies.

Exercise 8.10

(Moderate)

Using the general formula $\Delta x = \sqrt{\hbar/2m\omega}$, verify the zero-point spread quoted in the worked example for the 1.0 kg, $k = 100$ N/m macroscopic oscillator, and separately for a hydrogen molecule ($\mu \approx 8.4 \times 10^{-28}$ kg, $k \approx 570$ N/m), comparing your result to the H–H bond length of about 74 pm.

Exercise 8.11

(Moderate)

Explain, using the uncertainty principle rather than solving the Schrödinger equation directly, why both the infinite square well and the harmonic oscillator must have a ground-state energy strictly greater than the classical minimum ($E = 0$ in both cases), and why this argument would not apply to a classical (macroscopic) oscillator or box.

Exercise 8.12

(Moderate)

A hydrogen molecule vibrates with quantized energy levels $E_n = (n + \tfrac{1}{2})\hbar\omega$. If the molecule is initially in the $n = 2$ state, list the possible photon energies (in terms of $\hbar\omega$) it could emit in a single transition consistent with the selection rule $\Delta n = \pm 1$, and explain why a transition directly from $n = 2$ to $n = 0$ does not occur by single-photon emission.

Exercise 8.13
(Moderate)

Explain qualitatively, using the correspondence principle, why the probability density $|\psi_n(x)|^2$ of a highly excited ($n \gg 1$) harmonic-oscillator state should be largest near the classical turning points and smallest near $x = 0$ — the opposite of the ground-state ($n = 0$) distribution — and relate this to how much time a classical oscillator of the same energy spends near each of those locations.

Exercise 8.14
(Moderate)

A particle of energy E is incident on a potential step of height $V_0 > E$ (rather than $V_0 < E$ as in the worked example). (a) Explain why $R = 1$ exactly in this case, in terms of the number of available propagating modes on the far side of the step. (b) Despite total reflection, explain (referring to the finite-well discussion) why the probability density is not simply zero for $x > 0$.

Exercise 8.15
(Moderate)

A particle in an infinite square well of width L is prepared in the $n = 3$ state. (a) Sketch (in words) the shape of $|\psi_3(x)|^2$, stating the number and approximate locations of its nodes. (b) If the well width is doubled to $2L$ with the particle remaining in the state with the same quantum number $n = 3$, by what factor does E_3 change?

Exercise 8.16
(Moderate)

Explain why a free-particle plane wave $\Psi \propto e^{i(kx-\omega t)}$ cannot be normalized, and explain, referring to the uncertainty principle, why this is an unavoidable consequence of the wave having a perfectly sharp, definite momentum $p = \hbar k$.

Exercise 8.17
(Challenging)

Using the even-solution matching condition $k\tan(kL/2) = \kappa$ for the symmetric finite square well, explain (without solving numerically) why increasing the well depth V_0 at fixed width L tends to increase the number of bound states, referring to how κ depends on V_0.

Exercise 8.18
(Challenging)

Repeat the alpha-decay order-of-magnitude estimate in the worked example, but for a barrier width of $L = 10$ fm instead of 7 fm, keeping $\Delta E = 20$ MeV unchanged. Find the new tunneling probability T and the resulting order-of-magnitude half-life estimate, and comment on how sensitively the result depends on the assumed barrier width.

Exercise 8.19
(Challenging)

State the number of nodes you expect for the harmonic-oscillator wave function $\psi_4(x)$ ($n = 4$), and compare this to the number of interior nodes of the $n = 5$ infinite-square-well wave function. Explain why the two specified states have the same number of nodes even though their quantum-number labels differ.

Exercise 8.20

(Challenging)

A macroscopic pendulum of mass 0.50 kg and angular frequency $\omega = 2.0$ rad/s is released from an amplitude corresponding to a total energy of 1.0×10^{-3} J. Estimate its effective quantum number n (from $E_n \approx n\hbar\omega$ for large n), and comment on why this pendulum's motion appears entirely classical and continuous despite formally obeying the same quantized-energy law as the harmonic oscillator of this chapter.

Chapter 9

Quantum Mechanics in Three Dimensions

Learning Objectives

By the end of this chapter, you should be able to:

- Write the time-independent Schrödinger equation in three dimensions and apply it to separable Cartesian potentials such as the 3D infinite box, including cases of partial and full degeneracy.
- Explain why central (spherically symmetric) potentials are naturally treated in spherical coordinates, and describe the separation of variables $\psi(r, \theta, \phi) = R(r)Y(\theta, \phi)$.
- Explain the origin and physical meaning of the centrifugal barrier term in the radial equation for a central potential.
- Identify the three quantum numbers (n_r, ℓ, m_ℓ) that arise from solving a central-potential problem and state the physical quantity each one labels; for the Coulomb potential, relate the radial quantum number to the principal quantum number n.
- State the quantization of orbital angular momentum magnitude and z-component, and explain the physical meaning of each, including the commutation relations that forbid simultaneous knowledge of more than one Cartesian component.
- Describe the isotropic three-dimensional harmonic oscillator as a second exactly solvable central potential and relate its degeneracy structure to the nuclear shell model.
- Explain why the three-dimensional treatment is the necessary foundation for the hydrogen atom, developed in Chapter 10.

Introduction

Chapter 8 solved the Schrödinger equation for several one-dimensional potentials, but every atom, molecule, and nucleus is a three-dimensional object, and the most important potential in atomic physics — the Coulomb attraction between an electron and a nucleus — depends only on the distance r from a fixed center, not on a single Cartesian coordinate. This chapter extends the Schrödinger equation to three dimensions and develops the machinery — separation of variables in spherical coordinates, the centrifugal barrier in the radial equation, and the angular momentum quantum numbers that emerge from it — needed to solve any **central-potential** problem, of which the hydrogen atom (Chapter 10) is the most important example. The results obtained here, especially the quantization of angular momentum, apply unchanged

to every central potential, not just the Coulomb potential, and reappear throughout atomic, molecular, and nuclear physics.

9.1 Three Dimensions and Central Potentials

The Schrödinger Equation in Three Dimensions

Generalizing Chapter 8 to three spatial dimensions, the time-independent Schrödinger equation for a particle of mass m in a potential $V(x, y, z)$ is

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}\right)+V(x,y,z)\,\psi=E\psi, \tag{9.1}$$

or, compactly, $-\dfrac{\hbar^2}{2m}\nabla^2\psi + V\psi = E\psi$, where ∇^2 is the Laplacian operator. As in one dimension, $|\psi(x,y,z)|^2\,dV$ gives the probability of finding the particle in the volume element dV about (x, y, z), and ψ must be normalizable: $\int |\psi|^2\,dV = 1$.

> **The Laplacian.** $\nabla^2 = \dfrac{\partial^2}{\partial x^2}+\dfrac{\partial^2}{\partial y^2}+\dfrac{\partial^2}{\partial z^2}$ is a single, coordinate-independent operator; later in this chapter it reappears written in spherical coordinates, where it looks more complicated but means exactly the same thing.

When the potential is **separable in Cartesian coordinates**, $V(x,y,z) = V_1(x) + V_2(y) + V_3(z)$, the equation can be solved by seeking product solutions $\psi(x,y,z) = X(x)Y(y)Z(z)$, which reduces the problem to three independent one-dimensional Schrödinger equations, one per coordinate, each solved exactly as in Chapter 8. For a **3D infinite box** of side lengths L_x, L_y, L_z (a direct generalization of the infinite square well), this gives

$$E_{n_x,n_y,n_z} = \frac{h^2}{8m}\left(\frac{n_x^2}{L_x^2}+\frac{n_y^2}{L_y^2}+\frac{n_z^2}{L_z^2}\right), \qquad n_x, n_y, n_z = 1, 2, 3, \ldots, \tag{9.2}$$

with three independent quantum numbers, one per dimension. A notable feature not seen in one dimension: for a cubic box ($L_x = L_y = L_z = L$), distinct combinations of (n_x, n_y, n_z) — e.g., $(2,1,1)$, $(1,2,1)$, $(1,1,2)$ — can give the *same* total energy. This is called **degeneracy**, and it is a recurring feature of higher-dimensional quantum systems, generally traceable to an underlying symmetry of the potential (here, the equivalence of the three directions in a cube).

> **Tip**
> Before grinding through a degeneracy count case by case, look at the potential's symmetry first. A cubic box ($L_x = L_y = L_z$) is invariant under permuting x, y, z, so energies depend only on the *set* $\{n_x, n_y, n_z\}$, and permutations of a given triple are automatically degenerate — no need to check them one at a time. If only two sides are equal, only swaps of those two are guaranteed to be degenerate; if all three sides differ, don't expect any exact degeneracies at all beyond numerical coincidence.

Worked Example: A Rectangular Quantum Dot

A semiconductor quantum dot confines an electron in a box with $L_x = L_y = 5.0$ nm but $L_z = 10.0$ nm (a "square" cross-section, but elongated along z) — a shape with less symmetry than a cube, but more than a fully generic box. Using $E_n(\text{eV}) = 0.376\ \text{eV}\cdot\text{nm}^2 \times n^2/L^2(\text{nm})$ (a convenient rewriting of Equation (9.2) for an electron), the ground state has

$$E_{1,1,1} = (0.376\,\text{eV}\cdot\text{nm}^2)\left[\frac{1}{(5.0\ \text{nm})^2} + \frac{1}{(5.0\ \text{nm})^2} + \frac{1}{(10.0\ \text{nm})^2}\right] = (0.376\,\text{eV}\cdot\text{nm}^2)(0.090\,\text{nm}^{-2}) \approx 34\,\text{meV}, \tag{9.3}$$

a confinement energy of a few tens of millielectronvolts, comparable to k_BT at room temperature (≈ 26 meV) — which is why quantum-dot devices are often cooled to enhance the visibility of their discrete energy levels. Because $L_x = L_y \neq L_z$, the box retains a *partial* symmetry: states (n_x, n_y, n_z) and (n_y, n_x, n_z) are degenerate for any choice of n_z (swapping the two equal dimensions changes nothing), but states that instead permute the unequal dimension, such as $(1, 1, 2)$ versus $(2, 1, 1)$, are generally *not* degenerate with $(1, 1, 1)$ or with each other. A box with $L_x = L_y = L_z$ would restore full degeneracy among all permutations of a given (n_x, n_y, n_z) triple, while a fully generic box ($L_x \neq L_y \neq L_z$) would show no degeneracy at all — the amount of degeneracy is a direct fingerprint of how much geometric symmetry the confining potential actually has.

Central Potentials and Spherical Coordinates

The Coulomb potential, and most potentials of physical interest in atomic and nuclear physics, depend only on the distance from a fixed point: $V(x, y, z) = V(r)$, where $r = \sqrt{x^2 + y^2 + z^2}$. Such a potential is generally not separable in Cartesian coordinates (the isotropic oscillator discussed later is an important exception), but it *is* separable in **spherical coordinates** (r, θ, ϕ), precisely because its symmetry matches that coordinate system. Written in spherical coordinates, the Schrödinger equation for a central potential admits solutions of the separable form

$$\psi(r, \theta, \phi) = R(r)\, Y(\theta, \phi), \tag{9.4}$$

where $R(r)$, the **radial wave function**, depends on the specific form of $V(r)$ and carries the information about the particle's radial probability distribution, while $Y(\theta, \phi)$, the **angular wave function**, turns out to be *completely independent of the specific form of* $V(r)$ — it is determined entirely by the requirement that ψ be single-valued and well-behaved on the sphere, and is therefore the same set of functions for the hydrogen atom, a 3D harmonic oscillator, or any other central potential.

Spherical Harmonics: The Lowest Few Explicitly

The angular functions $Y(\theta, \phi)$ that solve the angular equation are called **spherical harmonics**, conventionally written $Y_\ell^{m_\ell}(\theta, \phi)$ and labeled by exactly the two angular quantum numbers introduced below. The lowest few, up to overall normalization constants, are

$$Y_0^0 = \text{constant}, \qquad Y_1^0 \propto \cos\theta, \qquad Y_1^{\pm 1} \propto \sin\theta\, e^{\pm i\phi}, \tag{9.5}$$

and already illustrate the general pattern. The $\ell = 0$ harmonic, Y_0^0, is completely independent of θ and ϕ — an s-state wave function is **spherically symmetric**, with a probability

density depending only on r, the same in every direction from the center. The $\ell = 1$ harmonics depend on angle: Y_1^0, proportional to $\cos\theta$, is largest along the $\pm z$-axis and vanishes in the xy-plane (a p_z-type angular distribution, in the language used for atomic orbitals in Chapter 10), while $Y_1^{\pm 1}$, proportional to $\sin\theta\, e^{\pm i\phi}$, is largest in the xy-plane and vanishes along the z-axis. Each higher ℓ introduces additional angular structure — more lobes, more angular nodes — but, crucially, the same three functions $Y_1^0, Y_1^{+1}, Y_1^{-1}$ describe the angular dependence of a p-state electron in hydrogen, a p-state neutron in a nuclear shell-model potential, or a p-state particle in the isotropic oscillator discussed later in this chapter, since (as already emphasized) the angular equation and its solutions never reference $V(r)$ at all.

Counting nodes. A real orbital with angular quantum number ℓ has ℓ angular nodes in total. For the complex spherical harmonic $Y_\ell^{m_\ell}$ itself, the associated Legendre factor gives $\ell - |m_\ell|$ polar nodes; the phase factor $e^{im_\ell\phi}$ does not create additional zeros. Real combinations such as p_x, p_y, and p_z redistribute the nodes into familiar nodal planes — the same node-counting logic used for 1D bound states in Chapter 8.

The Radial Equation and the Centrifugal Barrier

Substituting $\psi = R(r)Y(\theta,\phi)$ into the full Schrödinger equation and separating variables reduces the radial part to an ordinary differential equation for $R(r)$. It is standard, and illuminating, to write this in terms of $u(r) \equiv rR(r)$, in which case the radial equation takes a form directly analogous to a *one-dimensional* Schrödinger equation for $u(r)$ on the half-line $r > 0$:

$$-\frac{\hbar^2}{2m}\frac{d^2u}{dr^2} + \left[V(r) + \frac{\hbar^2\,\ell(\ell+1)}{2mr^2}\right]u = Eu, \tag{9.6}$$

where ℓ is the orbital angular momentum quantum number introduced below. The extra term, $\hbar^2\ell(\ell+1)/2mr^2$, is called the **centrifugal barrier**: it behaves as an additional repulsive potential that grows without bound as $r \to 0$ (for any $\ell > 0$), pushing the particle's radial probability density away from the origin. This is the quantum-mechanical counterpart of the classical fact that a particle with nonzero angular momentum cannot reach the force center while its angular momentum remains nonzero. For $\ell = 0$ (an **s-state**, in the spectroscopic notation below), the centrifugal barrier vanishes entirely, and for the usual regular central potentials only $\ell = 0$ states can have a nonvanishing probability density exactly at $r = 0$ — a fact used directly in Chapter 10 to explain, for instance, which hydrogen atomic states can undergo processes that require the electron to overlap with the nucleus.

Deriving the Radial Equation for u(r) = rR(r)

Separating $\psi = R(r)Y(\theta,\phi)$ in the full Schrödinger equation leaves $R(r)$ obeying

$$-\frac{\hbar^2}{2m}\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \left[V(r) + \frac{\hbar^2\ell(\ell+1)}{2mr^2}\right]R = ER, \tag{9.7}$$

with the $\ell(\ell+1)$ eigenvalue supplied by the angular equation. The first term mixes a first and second derivative of R, which is exactly what the substitution $u(r) \equiv rR(r)$, i.e. $R = u/r$, is designed to clean up:

$$\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = \frac{d}{dr}\left[r^2\frac{d}{dr}\left(\frac{u}{r}\right)\right] = \frac{d}{dr}\left(r\frac{du}{dr} - u\right) = r\frac{d^2u}{dr^2}. \tag{9.8}$$

Multiplying the radial equation by r and substituting this result gives

$$-\frac{\hbar^2}{2m}\frac{d^2u}{dr^2} + \left[V(r) + \frac{\hbar^2\ell(\ell+1)}{2mr^2}\right]u = Eu, \tag{9.9}$$

exactly the one-dimensional-looking form quoted in the main text. The simplification comes with a boundary condition: for a usual regular central potential, physical regularity requires $R(r) = u(r)/r$ to stay finite as $r \to 0$, and hence $u(0) = 0$ — playing the same role here that $\psi(0) = 0$ plays at the rigid wall of a one-dimensional infinite well.

9.2 Orbital Angular Momentum

The separation above is not a mathematical accident: it reflects the fact that a central potential exerts no torque about the force center (the force is always radial), so **orbital angular momentum**, $\vec{L} = \vec{r} \times \vec{p}$, is conserved, exactly as in classical central-force motion (e.g., Kepler orbits). Solving the angular equation subject to the single-valuedness of $Y(\theta, \phi)$ shows that the magnitude and one Cartesian component (conventionally the z-component) of $\vec{L}$ are simultaneously quantized:

$$L = \sqrt{\ell(\ell+1)}\,\hbar, \qquad \ell = 0, 1, 2, \ldots, \tag{9.10}$$

$$L_z = m_\ell \hbar, \qquad m_\ell = -\ell, -\ell+1, \ldots, 0, \ldots, \ell-1, \ell, \tag{9.11}$$

where ℓ is the **orbital angular momentum quantum number** and m_ℓ is the **magnetic quantum number**, so named because L_z determines how the system's energy shifts in an external magnetic field (Chapter 11). For a given ℓ, there are $2\ell + 1$ allowed values of m_ℓ, corresponding to $2\ell + 1$ distinct orientations of the angular momentum vector relative to the chosen z-axis — a specific, testable manifestation of **space quantization**: the orbital angular momentum vector does not merely have a quantized *length*, it can only point in a discrete set of directions relative to an external axis, rather than any direction whatsoever as classical mechanics would allow.

> **Warning**
> The bound $\ell \leq n-1$ that holds for Coulomb bound states is easy to mistake for a universal law of quantum mechanics — it isn't. That particular relationship between n and ℓ is a special feature of the Coulomb potential, worked out in Chapter 10, where the radial equation happens to tie the two labels together. For a generic central potential, ℓ simply runs over $0, 1, 2, \ldots$ with no corresponding ceiling; the isotropic harmonic oscillator later in this chapter is a clean counterexample, where n_r and ℓ combine into $N = 2n_r + \ell$ rather than one bounding the other.

Two features are worth emphasizing, since both run against classical intuition. First, $L = \sqrt{\ell(\ell+1)}\hbar$, not $\ell\hbar$ — the "extra" factor means that for any $\ell > 0$ the angular momentum vector's length is slightly *larger* than its maximum possible z-component, $m_{\ell,\max}\hbar = \ell\hbar$; the vector can never point exactly along the z-axis. Second, because L_x and L_y are not simultaneously measurable with L_z, only the magnitude L and a single component L_z can be assigned definite values at once — the other two components remain fundamentally indeterminate, consistent with the vector never lying exactly along any single axis.

Historically, states of a given ℓ are labeled by spectroscopic letters inherited from early atomic spectroscopy: $\ell = 0, 1, 2, 3, 4, \ldots$ are denoted $s, p, d, f, g, \ldots$ respectively — a labeling convention used throughout atomic physics (Chapters 10–11) and retained today purely by tradition.

Why m_ℓ Is Called the Magnetic Quantum Number

The name "magnetic quantum number" is not arbitrary bookkeeping. A charged particle with orbital angular momentum $\vec{L}$ circulating about a center behaves, classically and quantum mechanically alike, as a tiny current loop with an associated **orbital magnetic dipole moment**,

$$\vec{\mu}_L = -\frac{e}{2m}\vec{L} \tag{9.12}$$

for an electron of charge $-e$, directed opposite to $\vec{L}$ because the electron's charge is negative. Placed in an external magnetic field $\vec{B}$ along the z-axis, this moment contributes an additional term to the electron's energy, $-\vec{\mu}_L \cdot \vec{B} \propto m_\ell$, proportional directly to the magnetic quantum number: each of the $2\ell + 1$ otherwise energy-degenerate orientations of $\vec{L}$ acquires a distinct energy shift once a magnetic field is applied, splitting a single spectral line into $2\ell + 1$ closely spaced components. This splitting — the (normal) **Zeeman effect** — was observed spectroscopically well before quantum mechanics existed and is precisely why m_ℓ earned its name: it is the quantum number that controls how atomic energy levels respond to an external magnetic field, a connection developed quantitatively in Chapter 10 and Chapter 11, alongside the further complication (electron spin) required to explain the full richness of observed atomic spectra in a magnetic field.

Same letter, two meanings. In $\vec{\mu}_L = -\frac{e}{2m}\vec{L}$, the m is the electron's *mass*; the magnetic quantum number m_ℓ, used throughout this chapter, is a completely different, dimensionless number specifying orientation. Watch for both letters sitting side by side in expressions like this one.

That a magnetic field sorts a beam into a discrete number of orientations — rather than the continuum a classical dipole would give — is the content of Figure 9.1. Firing spin-$\frac{1}{2}$ atoms through a single analyzer splits the beam in two; chaining analyzers at different angles shows that the measurement does not simply read a pre-existing orientation.

Note

When Otto Stern and Walther Gerlach ran this experiment in 1922, quantum spin had not yet been proposed — Uhlenbeck and Goudsmit would not put it forward until 1925. Stern and Gerlach interpreted their two discrete silver-atom deflections as direct evidence for the *orbital* space quantization derived above, and the result was celebrated for exactly that reason. It later became clear that a silver atom's outer electron is in an s-state ($\ell = 0$, hence zero orbital magnetic moment — see the problems at the end of this chapter); the splitting Stern and Gerlach actually observed comes entirely from electron spin, a genuinely new form of angular momentum with no classical analog, introduced properly in Chapter 10.

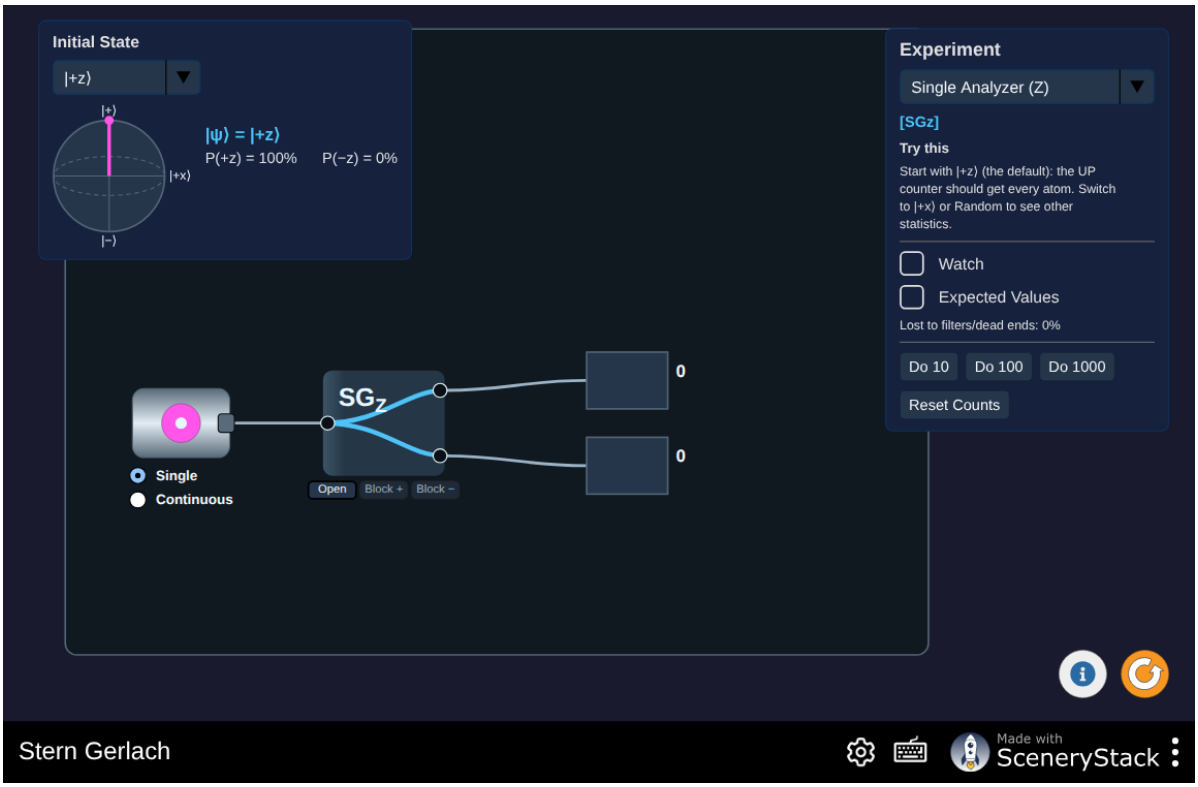

Figure 9.1: The Stern–Gerlach experiment, assembled from analyzers, magnets, and counters. Monte Carlo counts accumulate alongside the analytic quantum prediction, so the $2\ell + 1$ (or, for spin, $2s + 1$) discrete outcomes can be checked against the statistics. Interactive simulation: Stern Gerlach

Worked Example: The Vector Model for $\ell = 2$

An electron is in a state with orbital angular momentum quantum number $\ell = 2$ (a d-state). Its angular momentum magnitude is

$$L = \sqrt{\ell(\ell+1)}\,\hbar = \sqrt{6}\,\hbar \approx 2.449\,\hbar, \tag{9.13}$$

while the maximum possible z-component is $L_{z,\max} = m_{\ell,\max}\hbar = 2\hbar$. The smallest possible angle between $\vec{L}$ and the z-axis, achieved when m_ℓ takes its largest value, is

$$\cos\theta = \frac{L_z}{L} = \frac{2\hbar}{2.449\,\hbar} = 0.816 \quad \Longrightarrow \quad \theta \approx 35.3^\circ, \tag{9.14}$$

confirming explicitly that $\vec{L}$ can never point exactly along z (which would require $\theta = 0^\circ$) no matter how large m_ℓ is chosen for a given ℓ; the discrepancy $\ell\hbar$ versus $\sqrt{\ell(\ell+1)}\hbar$ shrinks only in a relative sense as ℓ grows large, another instance of the correspondence principle encountered in Chapter 8.

The cone that this worked example describes is usually drawn as a static picture, which makes the indeterminacy of L_x and L_y look like an admission of ignorance about a vector that is really sitting still somewhere on the cone. The classical system that gets the geometry right is a gyroscope, Figure 9.2: gravity applies a torque perpendicular to the spin angular momentum, and the response is not to tip the axis over but to walk it around a cone at fixed polar angle, with $|\vec{L}|$ and L_z both constant while L_x and L_y oscillate. Put a magnetic moment in a field instead of a top in gravity and the equation of motion is the same one; the quantum content is only that the cone's opening angle may take just $2\ell + 1$ values.

Angular Momentum Commutation Relations

The impossibility of simultaneously sharp values for more than one Cartesian component of $\vec{L}$ is not merely observed; it follows from the algebra of the angular momentum operators themselves, which satisfy the commutation relations

$$[\hat{L}_x, \hat{L}_y] = i\hbar\hat{L}_z, \qquad [\hat{L}_y, \hat{L}_z] = i\hbar\hat{L}_x, \qquad [\hat{L}_z, \hat{L}_x] = i\hbar\hat{L}_y \tag{9.15}$$

(a cyclic pattern in $x \to y \to z \to x$), together with $[\hat{L}^2, \hat{L}_z] = 0$. Two operators that do not commute cannot, in general, have simultaneous eigenstates — precisely the mathematical statement of the uncertainty principle applied to angular momentum, directly analogous to the position–momentum commutator underlying the Heisenberg relation of Chapter 7. Because $\hat{L}^2$ commutes with $\hat{L}_z$ (and likewise with $\hat{L}_x$ and $\hat{L}_y$), a quantum state can be prepared with simultaneously sharp values of L and one chosen component — exactly the states labeled by ℓ and m_ℓ above — but, for $\ell > 0$, never with sharp values of L, L_z, and L_x all at once.

> **See Also**
> The impossibility of jointly sharp L_x, L_y, L_z is the same phenomenon, in different clothing, as the position–momentum uncertainty relation built from $[\hat{x}, \hat{p}_x] = i\hbar$ in Chapter 7: in both cases it is the failure of two operators to commute that forbids any state from having simultaneously sharp values of the corresponding observables.

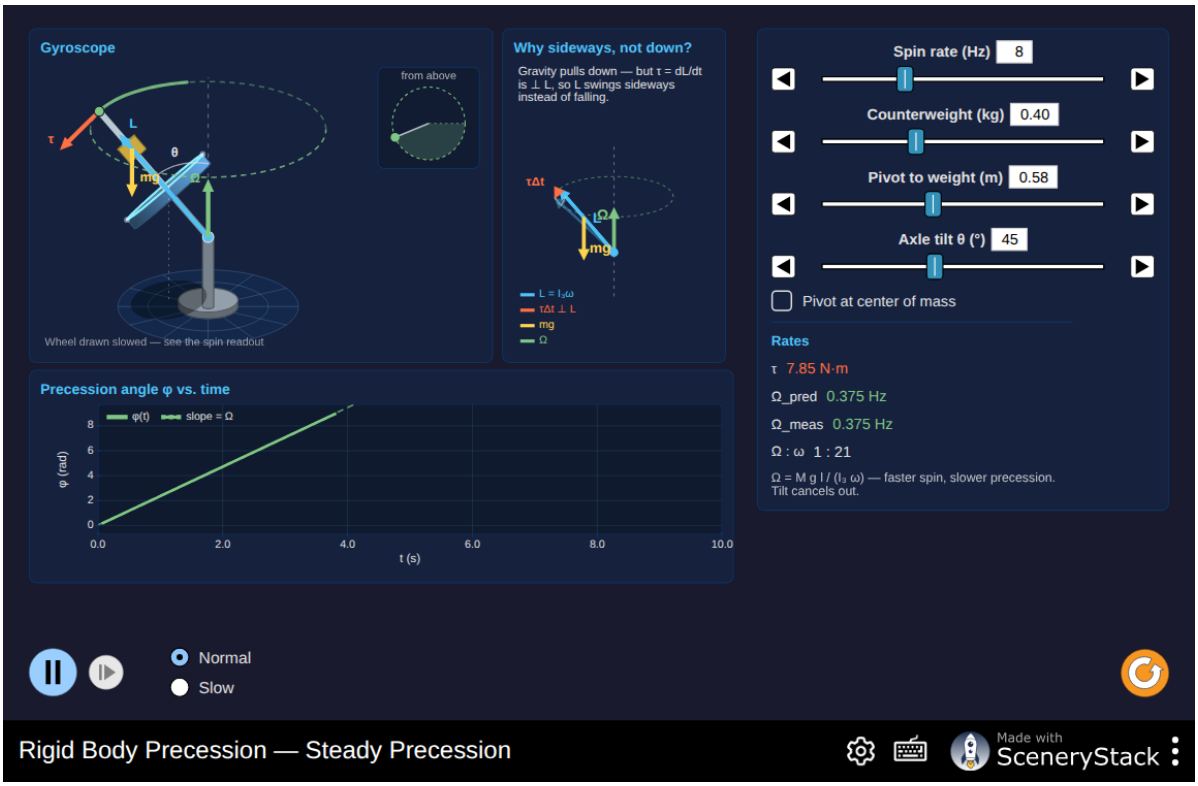

Figure 9.2: A spinning top under gravity. In steady precession the torque stays perpendicular to $\vec{L}$, so it changes the direction of the angular momentum without changing either its magnitude or its vertical component — the classical motion behind the vector model's cone, and behind the Larmor precession of a magnetic moment in a field (Chapter 10). Interactive simulation: Precession

9.3 Quantum Numbers and the Isotropic Oscillator

The Three Quantum Numbers of a Central-Potential Bound State

Solving the full three-dimensional problem for a bound state in a central potential $V(r)$ produces three quantum numbers, each arising from a separate boundary condition in the separation of variables:

- n_r, the **radial quantum number**, $n_r = 0, 1, 2, \ldots$, counting the radial nodes and arising from solving the radial equation subject to normalizability. The energy generally depends on both n_r and ℓ; for the Coulomb potential it is convenient to replace n_r and ℓ by the principal quantum number $n = n_r + \ell + 1$, as shown in Chapter 10;
- ℓ, the **orbital angular momentum quantum number**, $\ell = 0, 1, 2, \ldots$, governing the magnitude of orbital angular momentum;
- m_ℓ, the **magnetic quantum number**, $m_\ell = -\ell, \ldots, \ell$, governing the orientation of orbital angular momentum relative to a chosen axis.

This trio of labels, with the same angular ranges and physical meanings, appears in every central-potential problem in this book — it is a consequence of three-dimensional rotational symmetry, not a special feature of any one potential. Only the Coulomb problem promotes the combination $n_r + \ell + 1$ to a principal quantum number whose energy is independent of ℓ; Chapter 10 specializes the general machinery to that case.

A Second Exactly Solvable Central Potential: The Isotropic Harmonic Oscillator

The Coulomb potential of Chapter 10 is the most important central potential in atomic physics, but it is not the only one that can be solved exactly. The **isotropic three-dimensional harmonic oscillator**, $V(r) = \frac{1}{2}m\omega^2 r^2$, is a second example, and it is central (in the technical sense of depending only on r) even though it is also separable in Cartesian coordinates, since $r^2 = x^2 + y^2 + z^2$ splits into three independent one-dimensional oscillators. Solved in Cartesian form, using the one-dimensional harmonic-oscillator result of Chapter 8 three times over, the energy levels are

$$E_{n_x,n_y,n_z} = \left(n_x + n_y + n_z + \tfrac{3}{2}\right)\hbar\omega \equiv \left(N + \tfrac{3}{2}\right)\hbar\omega, \qquad N \equiv n_x + n_y + n_z = 0, 1, 2, \ldots, \tag{9.16}$$

depending only on the *sum* N, not on how it is distributed among n_x, n_y, n_z individually. The isotropic oscillator has continuous rotational symmetry, which guarantees the $2\ell+1$ degeneracy of each angular-momentum multiplet; its larger degeneracy across different ℓ values at fixed N reflects an additional oscillator symmetry. The same energy levels can equally well be labeled, via the spherical-coordinate separation of this chapter, by a radial quantum number and ℓ, with $N = 2n_r + \ell$; the two labeling schemes describe the same physical states, related by a change of basis, and the total degeneracy of a given level N works out to $(N+1)(N+2)/2$.

Where $(N+1)(N+2)/2$ comes from. This is the number of ways to write a fixed total N as an ordered sum of three non-negative integers, $n_x + n_y + n_z = N$ — a standard "stars and bars" counting result, independent of any physics.

This may look like a mathematical curiosity, but the isotropic harmonic oscillator is directly useful: it is the starting point for the nuclear **shell model** (Chapter 13), in which each

nucleon moves, to a first approximation, in an average central potential produced by all the other nucleons — a potential that resembles a finite well but is often approximated, for the purpose of a first, analytically tractable calculation, by an isotropic harmonic oscillator. The oscillator's degenerate energy levels, once a spin-orbit correction (introduced in Chapter 13) is added, reproduce the empirically observed nuclear **magic numbers** — proton or neutron counts (2, 8, 20, 28, 50, 82, 126) at which nuclei are unusually stable — in much the same way that filled electron shells explain the chemical stability of the noble gases (Chapter 11).

Comparing Degeneracy Across Three Central Potentials

It is worth pausing to compare the degeneracy patterns of the three central (or Cartesian-separable) potentials encountered so far, since the comparison sharpens exactly what "degeneracy reflects underlying symmetry" means in practice. The cubic infinite box has a *discrete* symmetry (the box is invariant only under swapping x, y, z and reflections, not under arbitrary rotations), and correspondingly modest degeneracies, arising only from specific numerical coincidences among sums of squares of integers. The isotropic harmonic oscillator has full continuous rotational symmetry plus an additional symmetry responsible for its larger degeneracy across different ℓ values at fixed $N = n_x + n_y + n_z$. The hydrogen atom's Coulomb potential (Chapter 10) has an additional degeneracy as well: not only is the energy independent of m_ℓ (true for *any* central potential, since no central potential singles out a preferred direction in space), but for the Coulomb potential specifically, the energy is also independent of ℓ for fixed principal n — an extra, "accidental" degeneracy not explained by rotational symmetry alone, and one of the most distinctive mathematical features of the $1/r$ potential, examined further in Chapter 10.

9.4 Summary

- The 3D Schrödinger equation, $-\dfrac{\hbar^2}{2m}\nabla^2\psi + V\psi = E\psi$, reduces to three independent 1D equations for a Cartesian-separable potential (e.g., the 3D infinite box), which can produce **degeneracy** — distinct quantum states sharing the same energy — as a signature of underlying symmetry; a box with partial symmetry (e.g., a square cross-section but different height) shows partial degeneracy, while a fully generic box shows none.
- A **central potential**, $V(r)$, is separable in spherical coordinates as $\psi(r,\theta,\phi) = R(r)Y(\theta,\phi)$; the angular part $Y(\theta,\phi)$ is universal, independent of the specific form of $V(r)$, while the radial equation for $u(r) = rR(r)$ contains a repulsive **centrifugal barrier** term $\hbar^2\ell(\ell+1)/2mr^2$ that vanishes only for $\ell = 0$.
- Orbital angular momentum is quantized in both magnitude, $L = \sqrt{\ell(\ell+1)}\hbar$ ($\ell = 0, 1, 2, \ldots$), and z-component, $L_z = m_\ell\hbar$ ($m_\ell = -\ell, \ldots, \ell$) — **space quantization** — with $2\ell+1$ allowed orientations for each ℓ. The commutation relations $[\hat{L}_x, \hat{L}_y] = i\hbar\hat{L}_z$ (and cyclic permutations) show algebraically why L_x and L_y remain simultaneously indeterminate once L and L_z are sharp.
- The **spherical harmonics** $Y_\ell^{m_\ell}(\theta,\phi)$ are the universal angular wave functions of any central potential; the $\ell = 0$ harmonic is spherically symmetric, while $\ell = 1$ harmonics have the angular shape familiar from p-orbitals in Chapter 10.
- Three quantum numbers, n_r, ℓ, m_ℓ, characterize a bound state in a central potential; for the Coulomb potential, the radial and angular labels are often reorganized as $n = n_r + \ell + 1$, ℓ, m_ℓ. States are conventionally labeled $s, p, d, f, \ldots$ for $\ell = 0, 1, 2, 3, \ldots$. The magnetic

quantum number m_ℓ controls how a state's energy shifts in an external magnetic field (the Zeeman effect), which is the origin of its name.

- The **isotropic 3D harmonic oscillator**, $V(r) = \frac{1}{2}m\omega^2 r^2$, is a second exactly solvable central potential, with energies $E_N = (N + \frac{3}{2})\hbar\omega$ depending only on $N = n_x + n_y + n_z = 2n_r + \ell$ and highly degenerate levels — the starting point for the nuclear shell model of Chapter 13.

9.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats.

1. **Multiple choice.** For orbital angular-momentum quantum number ℓ, how many values may $m\ell$ take? Choices: (A) $2\ell + 1$, from $-\ell$ through $+\ell$. (B) ℓ, from 0 through ℓ - 1. (C) 2ℓ, excluding zero. (D) Only one value, $m\ell = \ell$.

2. **True or false.** The centrifugal barrier in the radial Schrödinger equation vanishes for $\ell = 0$.

3. **Drag the words.** Complete the central-potential description. A central potential depends only on _____. Its wave function separates into _____. The universal angular functions are _____. The barrier term is proportional to _____. (Terms: r; radial and angular parts; spherical harmonics; $\ell(\ell + 1)/r^2$; θ only; Cartesian plane waves.)

4. **Fill in the blanks.** Complete the angular-momentum and oscillator counts. For $\ell = 2$ there are _____ allowed $m\ell$ values; oscillator triples with the same sum nx + ny + nz have the _____ energy.

5. **Mark the words.** Mark the pair of orbital-angular-momentum quantities that can be sharp simultaneously. "A central-potential state may have sharp L^2 and Lz. The noncommuting Cartesian components Lx and Ly cannot all be sharp at the same time."

9.6 Conceptual Questions

1. Why can a three-dimensional wave function be separated into radial and angular parts for a central potential?

2. Explain why L^2 and one component such as L_z can be known simultaneously, but all three Cartesian components cannot.

3. How does the centrifugal barrier change the probability of finding a particle near the origin?

4. What is the physical meaning of a spherical harmonic's nodal lines or surfaces?

5. Why does rotational symmetry lead naturally to energy degeneracy in m_ℓ?

6. How would breaking spherical symmetry with an external field affect that degeneracy?

9.7 Problems

Exercise 9.1
(Straightforward)

For a cubic 3D infinite box of side L, list the three lowest-lying distinct energy levels (in units of $h^2/8mL^2$) and the quantum-number triples (n_x, n_y, n_z) that produce each, noting any degeneracies.

Exercise 9.2
(Straightforward)

For the rectangular quantum dot of the worked example ($L_x = L_y = 5.0\,\text{nm}$, $L_z = 10.0\,\text{nm}$), find the energy (in meV) of the state $(n_x, n_y, n_z) = (1, 1, 2)$, and determine whether it is degenerate with any other low-lying state, explaining your reasoning from the box's symmetry.

Exercise 9.3
(Straightforward)

An electron is in a state with $\ell = 2$. (a) List all allowed values of m_ℓ. (b) Compute the magnitude L of its orbital angular momentum (in units of $\hbar$). (c) Compute the maximum possible value of L_z, and show it is strictly less than L, explaining why physically.

Exercise 9.4
(Straightforward)

How many distinct (ℓ, m_ℓ) combinations are allowed in the hydrogen-like (Coulomb) shell with principal quantum number $n = 3$? List them, grouped by ℓ, and give the spectroscopic letter for each ℓ value.

Exercise 9.5
(Straightforward)

Explain, using the uncertainty relation among the components of angular momentum, why an electron in a state of definite L and L_z cannot simultaneously have a definite value of L_x, and why this is consistent with the angular momentum vector never lying exactly along the z-axis.

Exercise 9.6
(Straightforward)

A particle is in a hydrogen-like (Coulomb) bound state with principal quantum number $n = 4$. What is the maximum possible orbital angular momentum quantum number ℓ it can have, and how many total (ℓ, m_ℓ) states are available in that n shell (summed over all allowed ℓ)?

Exercise 9.7
(Straightforward)

Explain qualitatively why the *angular* part of the wave function, $Y(\theta, \phi)$, does not depend on the specific functional form of $V(r)$, while the *radial* part, $R(r)$, does — referring to which term(s) in the separated Schrödinger equation involve $V(r)$ and which involve only the angular derivatives.

Exercise 9.8
(Moderate)

Explain, using the centrifugal barrier term $\hbar^2\ell(\ell+1)/2mr^2$, why an s-state ($\ell = 0$) electron can have a nonzero probability density at the nucleus ($r = 0$), while a p-state ($\ell = 1$) electron cannot — a fact revisited in Chapter 11's discussion of why s-electrons are especially effective at "penetrating" toward the nucleus in multi-electron atoms.

Exercise 9.9
(Moderate)

For an electron in a state with $\ell = 3$ (f-state), find (using the method of the worked example) the smallest possible angle between $\vec{L}$ and the z-axis, and compare it to the $\ell = 2$ result found in the text, commenting on the trend as ℓ increases.

Exercise 9.10
(Moderate)

Verify the commutation relation $[\hat{L}_z, \hat{L}_x] = i\hbar\hat{L}_y$ is consistent (in terms of index cycling $x \to y \to z \to x$) with the relation $[\hat{L}_x, \hat{L}_y] = i\hbar\hat{L}_z$ already stated, by writing out the analogous relation obtained by cycling indices twice, and state which single relation among $\hat{L}^2$, $\hat{L}_x$, $\hat{L}_y$, $\hat{L}_z$ guarantees that L and one Cartesian component can be simultaneously sharp.

Exercise 9.11
(Moderate)

For the isotropic three-dimensional harmonic oscillator, list all the Cartesian quantum-number triples (n_x, n_y, n_z) that give $N = n_x + n_y + n_z = 2$, verify there are six of them, and compare this total to the formula $(N + 1)(N + 2)/2$ quoted in the text.

Exercise 9.12
(Moderate)

Explain, in terms of symmetry, why the isotropic harmonic oscillator shows a higher degree of degeneracy at a given energy than the cubic infinite box does at a comparable energy level, even though both potentials are invariant under the same set of coordinate permutations ($x \leftrightarrow y \leftrightarrow z$).

Exercise 9.13
(Moderate)

The nuclear shell model uses an isotropic-oscillator-like potential to explain why nuclei with certain "magic" numbers of protons or neutrons (2, 8, 20, ...) are unusually stable. Using only the degeneracy formula $(N+1)(N+2)/2$ (ignoring, for this problem, the spin-orbit correction mentioned in the text), compute the cumulative number of single-particle states available after filling oscillator shells $N = 0$, $N = 1$, and $N = 2$ (counting each spatial state as available to 2 nucleons of a given type, for spin), and compare the resulting cumulative totals to the first few magic numbers.

Exercise 9.14
(Moderate)

A particle in a hydrogen-like (Coulomb) potential is in a state with principal quantum number $n = 2$. (a) List the allowed (ℓ, m_ℓ) combinations. (b) If a measurement of L_z yields the maximum possible value for the largest allowed ℓ at this n, state the values of ℓ and m_ℓ, and compute L and the angle between $\vec{L}$ and the z-axis.

Exercise 9.15
(Moderate)

An electron in a hydrogen-like atom is in a p-state ($\ell = 1$, $m_\ell = 0$), whose angular wave function is $Y_1^0 \propto \cos\theta$. (a) At what polar angle(s) θ does the angular probability density $|Y_1^0|^2$ vanish? (b) Explain, in terms of your answer, why this state is often called a "p_z" state, and contrast its angular shape with the spherically symmetric $\ell = 0$ (s-state) probability density.

Exercise 9.16
(Challenging)

An atom in a state with $\ell = 1$ is placed in an external magnetic field along the z-axis. Using the fact that the energy shift is proportional to m_ℓ, state how many distinct energy levels the original (field-free) $\ell = 1$ level splits into, and explain why an s-state ($\ell = 0$) shows no such splitting from orbital angular momentum alone.

Exercise 9.17
(Challenging)

Verify, by explicit substitution of $n_x = n_y = n_z = 0$ (so $N = 0$), that the ground state of the isotropic three-dimensional harmonic oscillator has energy $E_0 = \frac{3}{2}\hbar\omega$, and explain why this is exactly three times the one-dimensional zero-point energy $\frac{1}{2}\hbar\omega$ found in Chapter 8.

Exercise 9.18
(Challenging)

Explain, referring to the "accidental" degeneracy discussed in the text, what would have to be true of a central potential $V(r)$ for its energy levels to depend separately on n_r and ℓ — i.e., for the extra Coulomb-specific degeneracy to be absent — and state whether you would expect the isotropic harmonic oscillator (whose levels depend on $N = 2n_r + \ell$, not on n_r and ℓ separately) to show this same kind of extra degeneracy.

Exercise 9.19
(Challenging)

A beam of silver atoms (used in the Stern–Gerlach experiment discussed further in Chapter 10) has its outer electron in an s-state ($\ell = 0$). Using the magnetic-moment relation $\vec{\mu}_L = -(e/2m)\vec{L}$, explain why this electron's *orbital* angular momentum cannot be responsible for any splitting of the beam in an inhomogeneous magnetic field, foreshadowing the need for an additional angular-momentum-like degree of freedom (electron spin) introduced in Chapter 10.

Part IV

Atoms and Molecules

Chapter 10

The Hydrogen Atom

Learning Objectives

By the end of this chapter, you should be able to:

- Substitute the Coulomb potential into the general radial equation of Chapter 9 to obtain the resulting energy quantization for hydrogen, and compare it to the Bohr model's prediction.
- Explain, from the normalizability of the radial equation's solutions, why the principal quantum number n is discrete and why ℓ is restricted to $0, \ldots, n-1$.
- Write down the explicit hydrogen radial wave functions R_{10}, R_{20}, R_{21}, describe the resulting radial probability distributions, and use the effective (centrifugal) potential to explain why low-ℓ states penetrate closer to the nucleus.
- Extend the hydrogen solution to hydrogenic ions (e.g., He^{+}, Li^{2+}) and compute the Z^2 scaling of energy and $1/Z$ scaling of orbital size.
- Read an energy-level diagram for hydrogen and compute the wavelengths of lines in the Lyman, Balmer, and Paschen series.
- Enumerate the allowed quantum states for a given n and explain the origin of degeneracy in hydrogen.
- Explain electron spin and the Stern–Gerlach experiment that revealed it, and compute the force and deflection produced by a Stern–Gerlach magnet.
- Specify an electron's orbital and spin orientation using the magnetic quantum numbers m_ℓ and m_s, and compute the corresponding orbital and spin magnetic moments.
- Compute the frequency and wavelength shift produced by the (normal) Zeeman effect, and estimate the order of magnitude of fine-structure splitting in terms of the fine-structure constant α.
- Apply selection rules to determine which transitions between hydrogen energy levels are allowed.

Introduction

Chapter 9 developed the general machinery for any central potential: separation into radial and angular parts, and the universal quantization of orbital angular momentum. This chapter specializes that machinery to the single most important central potential in atomic physics — the Coulomb attraction between an electron and a proton — and solves it for the hydrogen atom, the only atom for which the Schrödinger equation can be solved exactly in closed form. The result reproduces (and explains, rather than assumes) the energy levels first found

empirically in atomic spectra and postulated ad hoc in the 1913 Bohr model, while revealing a far richer structure — angular momentum, spatial probability distributions, and a fourth quantum number, electron spin, with no classical counterpart at all.

10.1 Energy Levels and Hydrogenic Structure

The Radial Equation and Energy Quantization

For the hydrogen atom, the central potential is the Coulomb attraction between the electron (charge $-e$) and the proton (charge $+e$),

$$V(r) = -\frac{e^2}{4\pi\epsilon_0\, r}. \tag{10.1}$$

Substituting this $V(r)$ into the radial equation obtained from the separation $\psi = R(r)Y(\theta,\phi)$ (Chapter 9), and requiring $R(r)$ to be normalizable (i.e., to decay rather than blow up as $r \to \infty$, and to remain finite at $r = 0$), restricts the allowed energies to exactly the same discrete set found by Bohr in 1913 from an ad hoc semiclassical model:

$$E_n = -\frac{m_e e^4}{8\epsilon_0^2 h^2}\,\frac{1}{n^2} = -\frac{13.6\text{ eV}}{n^2}, \qquad n = 1, 2, 3, \ldots. \tag{10.2}$$

Reduced mass. E_n above technically uses the electron mass m_e alone, an approximation that treats the proton as infinitely heavy and fixed. The exact two-body treatment replaces m_e everywhere with the **reduced mass** $\mu = m_e m_p/(m_e + m_p) \approx 0.9995\, m_e$, shifting every hydrogen level by about 0.05% — small, but (see below) not unmeasurable.

This agreement is a triumph for the Schrödinger equation — it reproduces a result that matched atomic spectroscopy to remarkable precision — but the derivation and interpretation are entirely different from Bohr's. Bohr postulated that the electron moves on definite circular orbits, with angular momentum quantized as $L = n\hbar$ by assumption. The Schrödinger treatment makes no such assumption about orbits at all; instead, n emerges purely as an index counting the normalizable solutions of the radial equation, the electron has no well-defined trajectory, and — as shown below — the ground state ($n = 1$) actually has **zero** orbital angular momentum, in direct contradiction to Bohr's $L = n\hbar$. The numerical agreement in E_n is, in this sense, a coincidence specific to the particular form of the $1/r$ Coulomb potential, not a sign that the Bohr picture was substantially correct.

Note

The reduced-mass correction is not just bookkeeping. In 1931–32, Harold Urey searched for a heavier isotope of hydrogen by looking for a faint companion line shifted slightly to the blue of each ordinary hydrogen line — exactly the shift predicted because a heavier nucleus pulls μ closer to m_e and makes $|E_n|$ very slightly larger. Finding that predicted companion in a distilled hydrogen sample was the discovery of **deuterium**, and earned Urey the 1934 Nobel Prize in Chemistry — a case of a correction easy to dismiss as negligible turning out to carry real, measurable physics.

The mechanism behind the discreteness is worth making explicit, since the same pattern recurs throughout quantum mechanics. Solving the radial equation for a trial (negative) energy E produces a function that falls off as $e^{-\kappa r}$ at large r, with $\kappa = \sqrt{-2m_e E}/\hbar$, multiplied by an infinite series in powers of r; for almost any choice of E, that series itself grows fast enough that the full product still diverges as $r \to \infty$, exactly the unnormalizable blow-up that ruled out most trial energies in the finite square well of Chapter 8. Demanding a normalizable wave function forces the series to *terminate* after a finite number of terms — collapsing it from an infinite series into a finite polynomial — and this termination happens only when $1/\kappa a_0$ lands exactly on a positive integer, which is precisely the principal quantum number n. The same termination condition also caps the orbital quantum number: the polynomial's degree fixes a nonnegative integer $n_r = 0, 1, 2, \ldots$ (the number of interior radial nodes, discussed further below) related to n and ℓ by $n = n_r + \ell + 1$, so that $n_r \geq 0$ immediately forces $\ell \leq n - 1$ — and, for the ground state $n = 1$, forces $n_r = \ell = 0$ exactly, which is the direct algebraic reason the ground state must have zero orbital angular momentum, not merely an empirical fact to be taken on faith.

The distance between Bohr's picture and Schrödinger's is easiest to see by running them side by side, as in Figure 10.1. Both reproduce the same E_n and therefore the same emission lines, but only one of them places the electron on an orbit.

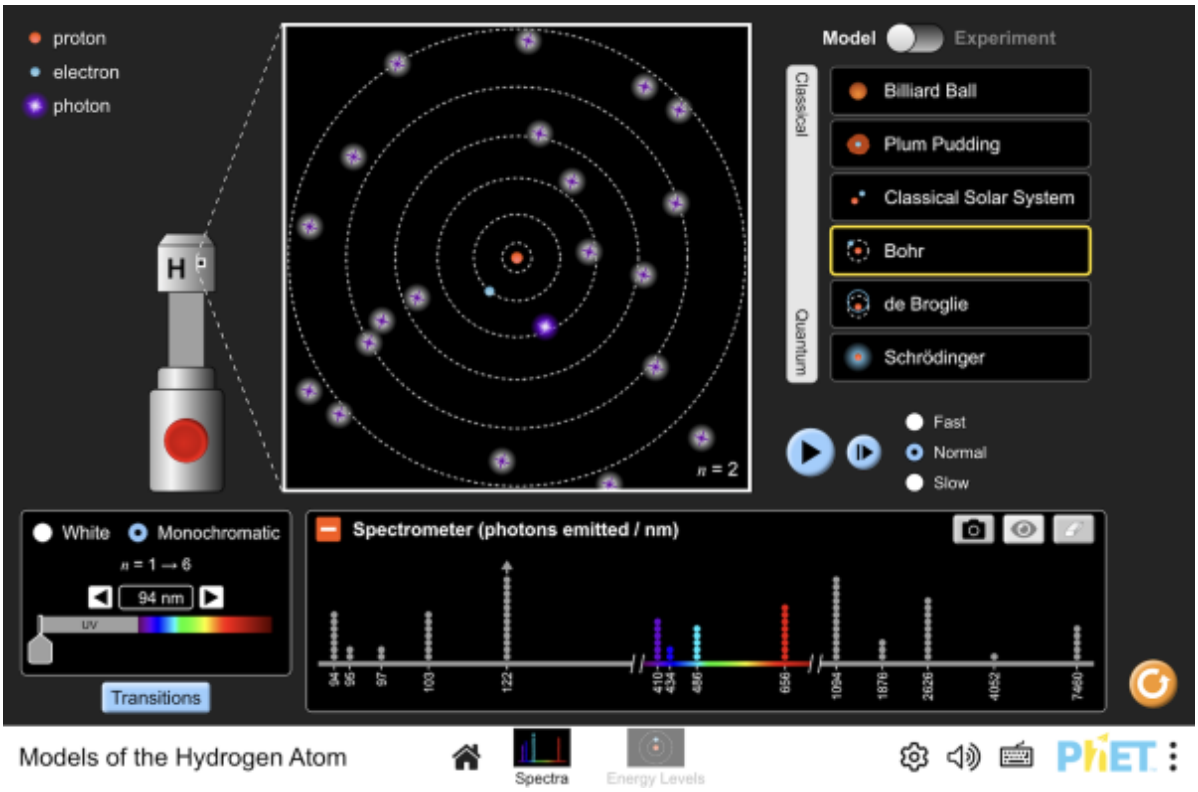

Figure 10.1: Successive models of the hydrogen atom — Bohr, de Broglie, Schrödinger, and their predecessors — each firing photons at an atom and producing a spectrum. The models that agree on E_n disagree entirely about where the electron is. Interactive simulation: Models of the Hydrogen Atom

Historical Context: Bohr's 1913 Model and Its Limits

It is worth pausing on why a model built on postulates that turned out to be simply wrong — definite orbits, angular momentum quantized as $L = n\hbar$ — nonetheless earned Niels Bohr the 1922 Nobel Prize and remains the picture most people first encounter. In 1913, thirteen years before Schrödinger's equation existed, Bohr combined classical circular-orbit mechanics with a single ad hoc quantization rule and reproduced the empirical Rydberg formula for hydrogen's spectral lines essentially exactly, including its dependence on nuclear charge for hydrogenic ions (below) — a stunning success for a model with no derivation behind its central assumption. The de Broglie standing-wave argument of Chapter 7 later supplied a retroactive justification for $L = n\hbar$, which is part of why the Bohr model survived as long as it did before being fully superseded.

> **See Also**
> The de Broglie justification for $L = n\hbar$ works out concretely: demanding that an integer number of de Broglie wavelengths, $\lambda = h/p$ (Chapter 7), fit exactly around a circular orbit of circumference $2\pi r$ gives $2\pi r = n\lambda = nh/p$, i.e. $L = rp = n\hbar$ — Bohr's quantization rule, recovered after the fact from wave interference rather than assumed outright.

The model's failures, however, were just as instructive as its success. Extending Bohr's orbit-quantization scheme to **helium** — even with the refinements (elliptical orbits, relativistic orbit precession) added by Arnold Sommerfeld through the 1910s — never produced a correct ionization energy or a stable ground-state configuration for a two-electron atom; the semiclassical machinery simply had no consistent way to handle two mutually interacting orbiting electrons. The model also could not predict *which* transitions between levels actually occur (the selection rules developed later in this chapter), could not account for the relative brightness of spectral lines, and — as already emphasized above — gets the ground state's angular momentum flatly wrong: Bohr's $n = 1$ orbit carries $L = \hbar$, while the true ground state has $L = 0$ and no orbit, well-defined trajectory, or definite radius at all, only the probability cloud shown in Figure 10.1. These failures, especially the inability to extend the model consistently beyond hydrogen, were a central motivation for the fully quantum-mechanical treatment developed in Chapters 8–9 and specialized to hydrogen in this chapter — a treatment that, unlike Bohr's, generalizes cleanly to helium and every other atom in Chapter 11.

Hydrogenic Ions: Scaling with Nuclear Charge

Everything derived above specializes the general central-potential machinery of Chapter 9 to a nuclear charge of exactly $+e$. The same radial equation applies unchanged, with only $e^2 \to Ze^2$ in the Coulomb potential, to any **hydrogenic (hydrogen-like) ion**: a single electron bound to a nucleus of charge $+Ze$, such as singly ionized helium He^+ ($Z = 2$) or doubly ionized lithium Li^{2+} ($Z = 3$). Repeating the normalizability argument above with the rescaled potential gives

$$E_n(Z) = -Z^2 \frac{13.6\ \text{eV}}{n^2}, \qquad a_n(Z) = \frac{n^2}{Z}\, a_0, \tag{10.3}$$

where $a_n(Z)$ is the characteristic Bohr length scale of the n-th shell. It is useful for comparing orbital sizes, but it is not the exact radial-probability peak for every (n, ℓ) state; those peak positions also depend on ℓ. Both dependences make direct physical sense: a larger nuclear charge attracts the electron more strongly, pulling its orbit *inward* (radius shrinks as



Figure 10.2: A step back to the atom itself: the proton–neutron–electron bookkeeping that both Bohr's orbits and the wave functions later in this chapter are trying to explain the behavior of. Watch on YouTube: What Are Atoms and Isotopes?

$1/Z$) while binding it *more tightly* — and because the potential energy itself scales as Z while the resulting spatial compression compounds that scaling once more, the binding energy grows as Z^2, not simply Z. This Z^2 scaling reappears with direct experimental consequence in Chapter 11, where it underlies Moseley's law for the energies of characteristic X-ray transitions in multi-electron atoms.

Worked Example: The Ionization Energy and Orbital Size of He^+

Find the ground-state ionization energy and the most probable electron–nucleus separation for He^+ ($Z = 2$), a one-electron ion, and compare both to hydrogen's.

$$E_1(\mathrm{He}^+) = -Z^2(13.6\ \mathrm{eV}) = -(2)^2(13.6\ \mathrm{eV}) = -54.4\ \mathrm{eV}, \tag{10.4}$$

so the ionization energy — the energy required to remove the electron entirely — is 54.4 eV, exactly four times hydrogen's 13.6 eV, matching the measured ionization energy of He^+. The orbital size scales the other way,

$$a_1(\mathrm{He}^+) = \frac{a_0}{Z} = \frac{0.0529\ \mathrm{nm}}{2} = 0.0265\ \mathrm{nm}, \tag{10.5}$$

half of hydrogen's Bohr radius: the electron in He^+ is bound four times more tightly and orbits, on average, twice as close to the nucleus as the electron in neutral hydrogen — both consequences of the same doubled nuclear charge, entering the energy quadratically and the size linearly (inversely).

Quantum Numbers and Degeneracy in Hydrogen

Solving the full three-dimensional problem gives states labeled by the same three quantum numbers introduced in Chapter 9 — n, ℓ, m_ℓ — but with a further restriction, specific to the $1/r$ Coulomb potential, tying ℓ to n:

$$n = 1, 2, 3, \ldots, \qquad \ell = 0, 1, \ldots, n-1, \qquad m_\ell = -\ell, \ldots, \ell. \tag{10.6}$$

Because the energy E_n depends *only* on n — not on ℓ or m_ℓ — every state sharing a given n is degenerate (equal in energy), regardless of its orbital angular momentum. Counting the total number of (ℓ, m_ℓ) combinations for a given n gives n^2 degenerate states (before accounting for electron spin, discussed below): for example, $n = 2$ admits $\ell = 0$ (one state, $m_\ell = 0$) and $\ell = 1$ (three states, $m_\ell = -1, 0, 1$), for $1 + 3 = 4 = 2^2$ states total. This ℓ-independence of the energy is itself notable — it does *not* hold for multi-electron atoms (Chapter 11), where the energy depends on ℓ as well as n, and is a special feature of the pure $1/r$ Coulomb potential (technically, a signature of a hidden extra symmetry, beyond ordinary rotational symmetry, unique to the $1/r$ potential).

The Hidden Symmetry Behind ℓ-Independent Degeneracy

Ordinary three-dimensional rotational symmetry guarantees that energy cannot depend on m_ℓ, since no direction in space is special — but it says nothing about ℓ. A generic central potential's energy levels *do* depend on ℓ, and indeed they do for every atom with more than one electron (Chapter 11). Hydrogen's extra, "accidental" ℓ-independence traces back to a conserved quantity unique to the exact $1/r$ potential: the classical **Laplace–Runge–Lenz vector**,

$$\vec{A} = \vec{p} \times \vec{L} - \frac{m_e e^2}{4\pi\epsilon_0}\,\hat{r}, \tag{10.7}$$

which points from the focus of a classical Kepler orbit toward its perihelion and is constant in time only for a force that falls off as *exactly* $1/r^2$ — not $1/r^{2.01}$, not $1/r^{1.99}$. Quantum mechanically, $\vec{A}$ becomes an operator that commutes with the Hamiltonian and connects states of different ℓ at the same n, enlarging the symmetry governing hydrogen from ordinary three-dimensional rotations, $SO(3)$, to a larger four-dimensional rotation group, $SO(4)$, whose representations turn out to be exactly the n^2-fold degenerate multiplets observed. This is the same underlying fact about the $1/r^2$ force law that makes classical Kepler orbits close on themselves without precessing — the classical and quantum "accidents" are one and the same.

10.2 Wave Functions, Spectra, and Spin

Wave Functions and Probability Distributions

The full wave functions, $\psi_{n\ell m_\ell}(r,\theta,\phi) = R_{n\ell}(r)\,Y_{\ell m_\ell}(\theta,\phi)$, have structure worth examining qualitatively even without their explicit algebraic form. The ground state, ψ_{100}, is spherically symmetric ($\ell = 0$, so Y_{00} is constant) and decays exponentially, $R_{10}(r) \propto e^{-r/a_0}$, where

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2} = 0.0529 \text{ nm} \tag{10.8}$$

is the **Bohr radius** — reappearing here not as the radius of a Bohr orbit but as the natural length scale over which the ground-state probability density falls off. The **radial probability distribution**, $P(r) = r^2|R_{n\ell}(r)|^2$ (the probability per unit r of finding the electron at distance r from the nucleus, obtained by integrating $|\psi|^2$ over all angles at fixed r), peaks at $r = a_0$ for the ground state — the most probable electron-nucleus distance in hydrogen's ground state is exactly the Bohr radius, even though the electron has zero orbital angular momentum and hence, unlike in the Bohr picture, is not "orbiting" in any classical sense.

> **Warning**
> "Most probable radius" is not the same thing as "average radius," and for hydrogen the two are not equal except by coincidence. The ground state's $P(r)$ peaks at $r = a_0$, but the curve falls off more gently on the outside of that peak than it rises on the inside, so the distribution is skewed toward larger r — enough that the average $\langle r\rangle = \int_0^\infty r P(r)\,dr$ works out to $\frac{3}{2}a_0$, fifty percent farther out than the peak. Don't assume a probability distribution is symmetric about its maximum just because it has a single, well-defined peak.

Written out in full, with proper normalization, the lowest few radial wave functions are

$$R_{10}(r) = \frac{2}{a_0^{3/2}}\,e^{-r/a_0}, \qquad R_{20}(r) = \frac{1}{2\sqrt{2}\,a_0^{3/2}}\left(2 - \frac{r}{a_0}\right)e^{-r/2a_0}, \qquad R_{21}(r) = \frac{1}{2\sqrt{6}\,a_0^{3/2}}\,\frac{r}{a_0}\,e^{-r/2a_0}, \tag{10.9}$$

each verified against the standard hydrogen wave-function table (e.g. OpenStax Vol. 3, §8.2). The pattern generalizes: every $R_{n\ell}(r)$ is an exponential e^{-r/na_0} (decay rate set only by n) multiplied by a polynomial in r of degree $n-1$, and the number of interior zeros of that

polynomial — the **radial nodes** counted below — is exactly $n_r = n - \ell - 1$, the same integer introduced above via $n = n_r + \ell + 1$. R_{10} ($n_r = 0$) is a bare exponential with no node; R_{20} ($n_r = 1$) has one node, at $r = 2a_0$, exactly where its parenthetical factor vanishes; R_{21} ($n_r = 0$, since its higher ℓ "uses up" the node budget at fixed $n = 2$) is a bare power of r times an exponential, with no node at all despite sharing $n = 2$ with R_{20}.

Figure 10.3 plots the resulting $P(r) = r^2|R_{n\ell}(r)|^2$ for six low-lying states, making the node-counting rule directly visible rather than merely asserted: each panel's dotted vertical lines mark its nodes, and the count in every panel matches $n - \ell - 1$ exactly — zero for $1s$, $2p$, and $3d$; one for $2s$ and $3p$; two for $3s$.

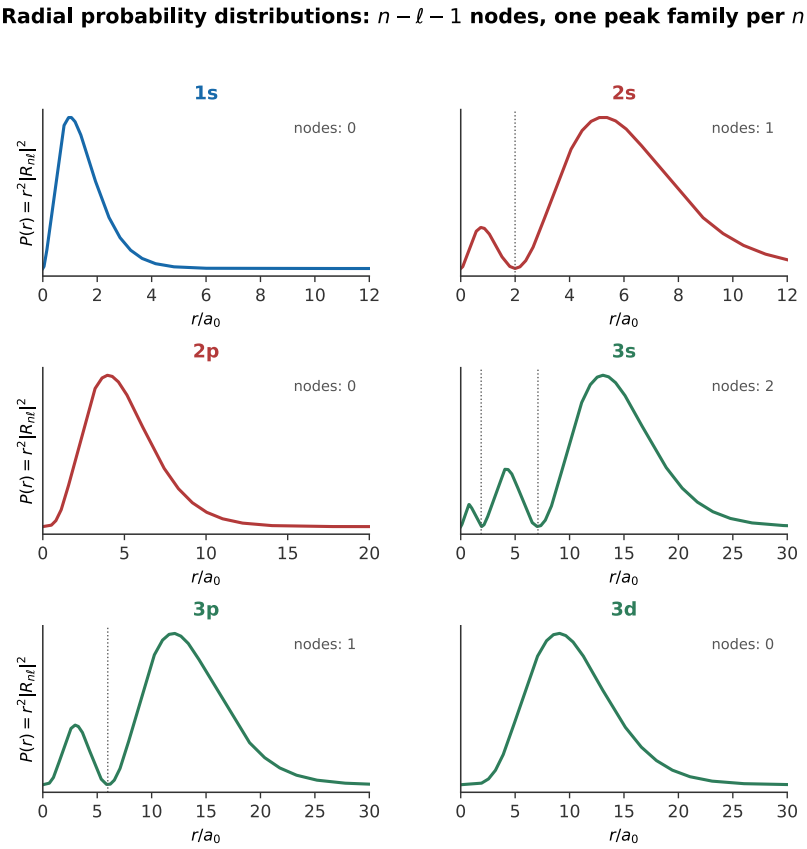

Figure 10.3: Radial probability distributions $P(r) = r^2|R_{n\ell}(r)|^2$ for the six lowest distinct (n, ℓ) combinations. Each state's node count matches $n - \ell - 1$ exactly, and every curve for a given n extends, on average, farther from the origin as n increases — the size of the atom really does grow with n, just not along a sharp Bohr orbit. Computed directly from the closed-form $R_{n\ell}(r)$ formulas; see `scripts/figures/`.

More generally, $R_{n\ell}(r)$ has $n - \ell - 1$ radial nodes (points, other than $r = 0$ and $r = \infty$, where the probability density vanishes), and the angular functions $Y_{\ell m_\ell}(\theta, \phi)$ have angular nodes (nodal planes or cones) whose count and shape depend on ℓ and m_ℓ — giving rise to the familiar s (spherical), p (dumbbell-shaped, with a single nodal plane through the origin), and d-orbital shapes used throughout chemistry (Chapter 12) to describe electron distributions in atoms and molecules.

Where the letters come from. s, p, d, f for $\ell = 0, 1, 2, 3$ are historical leftovers from nineteenth-century spectroscopy, abbreviating **s**harp, **p**rincipal, **d**iffuse, and **f**undamental — descriptions of how the corresponding spectral lines *looked*, decades before anyone knew what ℓ was. Past f, the letters simply continue alphabetically ($g, h, \ldots$).

The Effective Potential and the Centrifugal Barrier

Chapter 9 showed that the radial equation for any central potential can be recast as an effective one-dimensional Schrödinger equation for $u(r) = rR(r)$, governed by an **effective potential**

$$V_{\text{eff}}(r) = V(r) + \frac{\hbar^2\,\ell(\ell+1)}{2m_e r^2} = -\frac{e^2}{4\pi\epsilon_0 r} + \frac{\hbar^2\,\ell(\ell+1)}{2m_e r^2}, \tag{10.10}$$

the attractive Coulomb term plus the repulsive **centrifugal barrier** introduced there. For $\ell = 0$ (s-states), $V_{\text{eff}}(r) = V(r)$: nothing but the bare attractive well, all the way in to $r = 0$. This is why only $\ell = 0$ states can have a nonzero **probability density** $|\psi(0)|^2$ at the nucleus. The radial probability $P(r) = r^2|R(r)|^2$, however, is zero at $r = 0$ for every state because the spherical volume element contributes the factor r^2; accordingly, the $1s$, $2s$, and $3s$ curves in Figure 10.3 all rise from zero at the origin even though their wave functions themselves need not. For $\ell > 0$, the centrifugal term ($\propto 1/r^2$) dominates the attractive term ($\propto 1/r$) at small r and produces a genuine barrier that pushes the wave function away from the nucleus; the larger ℓ is, the taller that barrier and the farther out the first significant probability appears, exactly as seen in the figure, where $2p$, $3p$, and $3d$ all start at zero and rise more gradually than their same-n s-state counterparts, and where $3d$ — the largest ℓ at $n = 3$ — peaks farthest from the origin of the three $n = 3$ curves. This ℓ-dependent "penetration" toward the nucleus is purely qualitative for hydrogen's single electron, but it becomes quantitatively essential in Chapter 11, where it explains why an s-electron in a multi-electron atom is attracted more strongly by (and screens other electrons from) the nuclear charge than a p- or d-electron of the same n — the mechanism that breaks hydrogen's accidental ℓ-independent degeneracy once more than one electron is present.

Spectral Series and the Energy-Level Diagram

Every property established so far — discrete E_n, the selection rule developed below — can be assembled into a single picture: the ladder of allowed energies together with the transitions between them that are actually observed. Figure 10.4 draws that ladder for $n = 1$ through 5, together with the three most commonly tabulated spectral series, each named for whoever first cataloged it and each defined by a common *lower* level: the **Lyman series** ($n_f = 1$, entirely in the ultraviolet), the **Balmer series** ($n_f = 2$, spanning the visible), and the **Paschen series** ($n_f = 3$, infrared). Because the spacing between successive E_n shrinks rapidly as n grows (it falls off as roughly $1/n^3$ near the top of the ladder), the levels bunch up as they approach $E = 0$, and each series converges to a **series limit**: the shortest wavelength (highest photon energy) the series can produce, corresponding to a transition from $n_i \to \infty$ down to the series' fixed n_f, and equal to the energy $13.6\ \text{eV}/n_f^2$ needed to ionize the atom starting from level n_f. Beyond that limit the spectrum stops being a set of discrete lines at all, since a photoionized electron ($E > 0$) is no longer confined to a discrete spectrum and can carry away any leftover energy continuously.

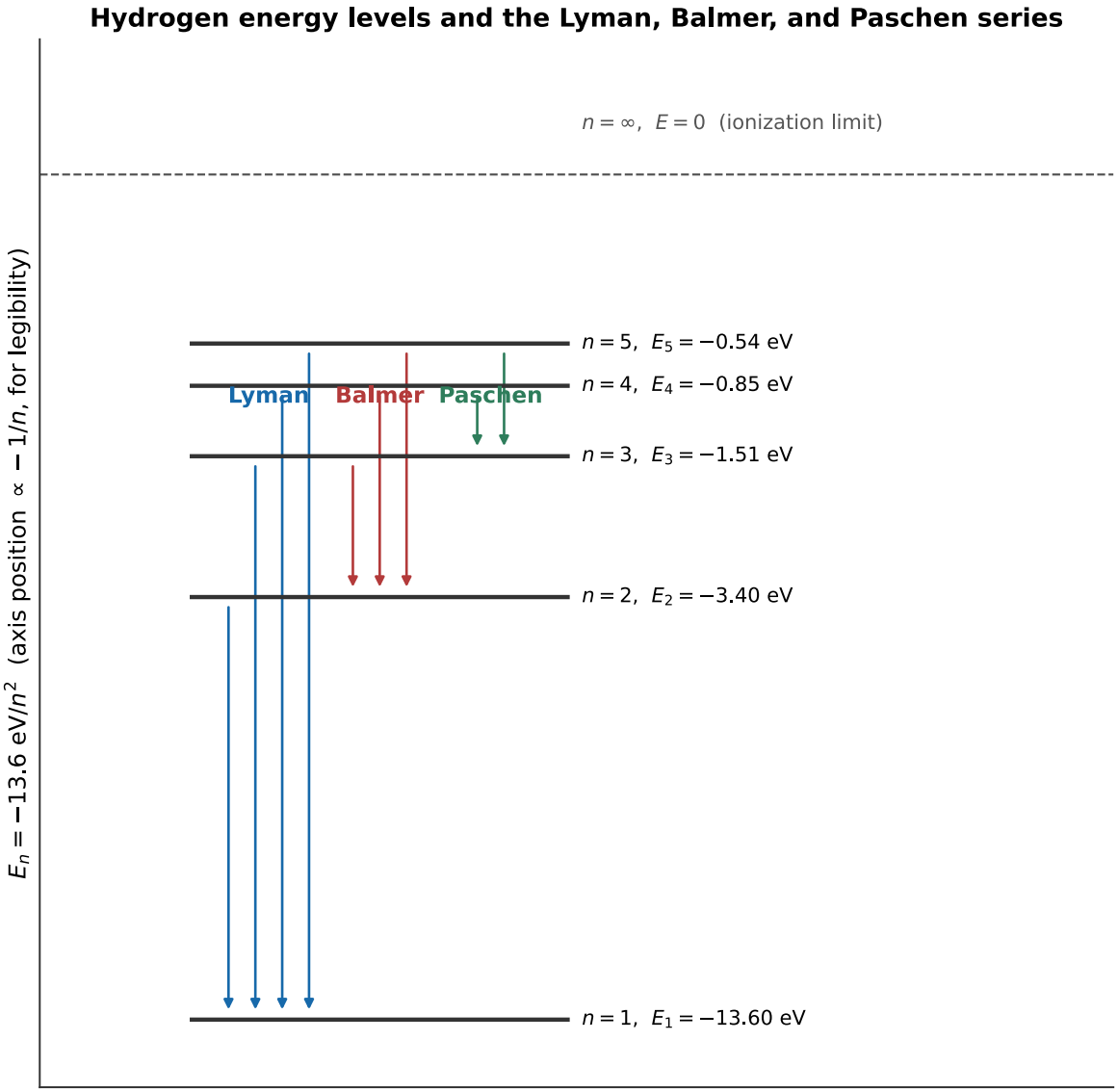

Figure 10.4: Hydrogen energy levels $n = 1$–5 (vertical position compressed for legibility; labels give the real $E_n = -13.6\ \mathrm{eV}/n^2$), with the Lyman, Balmer, and Paschen series drawn as clusters of downward transitions converging on $n_f = 1$, 2, and 3 respectively. Each series' shortest-wavelength member is its series limit, reached only as $n_i \to \infty$. Computed directly from E_n; see `scripts/figures/`.

Worked Example: The Balmer Series and the Visible Spectrum

The **Rydberg formula** is simply $E_n = -13.6\ \text{eV}/n^2$ rewritten as a wavelength via $hc/\lambda = |E_{n_i} - E_{n_f}|$:

$$\frac{1}{\lambda} = R\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right), \qquad R = 1.097 \times 10^7\ \text{m}^{-1}. \tag{10.11}$$

> **Two different R's.** The Rydberg constant R used here is unrelated to the radial wave function $R_{n\ell}(r)$ used earlier in this chapter — an unfortunate but standard notational collision. A bare R in a wavelength formula and an R carrying two subscripts and an r-dependence are always distinguishable from context.

Find (a) the wavelength of H_α, the first (longest-wavelength) line of the Balmer series ($n_i = 3 \to n_f = 2$), and (b) the Balmer series limit.

(a)

$$\frac{1}{\lambda} = (1.097 \times 10^7\ \text{m}^{-1})\left(\frac{1}{4} - \frac{1}{9}\right) = 1.524 \times 10^6\ \text{m}^{-1} \quad \Longrightarrow \quad \lambda = 656.3\ \text{nm}, \tag{10.12}$$

a deep red line — the H_α line used to image glowing hydrogen gas in nebulae and to trace the rotation of spiral galaxies.

(b) As $n_i \to \infty$, $1/n_i^2 \to 0$, so

$$\frac{1}{\lambda_{\text{limit}}} = \frac{R}{4} = 2.743 \times 10^6\ \text{m}^{-1} \quad \Longrightarrow \quad \lambda_{\text{limit}} = 364.6\ \text{nm}, \tag{10.13}$$

just past the violet edge of the visible spectrum, in the near ultraviolet. Every Balmer line falls between these two wavelengths, $364.6\ \text{nm} < \lambda \leq 656.3\ \text{nm}$ — which is why nearly the whole series, unlike the entirely ultraviolet Lyman series or the entirely infrared Paschen series, is visible to the human eye, and why Johann Balmer found the first four of these lines by curve-fitting alone in 1885, three decades before Bohr's model — or Schrödinger's equation — existed to explain them.

Electron Spin

By the mid-1920s, several pieces of spectroscopic evidence — most directly, the splitting of atomic beams passing through an inhomogeneous magnetic field — showed that the three quantum numbers (n, ℓ, m_ℓ) do not fully specify an electron's state. In the **Stern–Gerlach experiment** (1922), a beam of (electrically neutral) silver atoms was passed through a strongly inhomogeneous magnetic field and allowed to strike a detector screen. A classical magnetic dipole, oriented randomly, should be deflected by an amount depending continuously on its orientation, producing a single smeared-out band on the screen. Instead, the beam split into exactly **two** discrete spots, symmetric about the undeflected position — direct evidence of **space quantization** (Chapter 9) applied to a new, previously unsuspected degree of freedom, since the outermost electron in a silver atom happens to be in an s-state ($\ell = 0$, hence zero orbital angular momentum and no orbital magnetic moment to produce any deflection at all), so the observed splitting could not be due to orbital angular momentum.

The resolution, proposed by Samuel Goudsmit and George Uhlenbeck (1925), is that the electron possesses an intrinsic angular momentum, **spin**, $\vec{S}$, with no classical counterpart (it is not literally the electron spinning on its axis — such a picture leads to internal-consistency and speed-of-rotation contradictions and should be regarded purely as a suggestive name), quantized exactly as orbital angular momentum is, but with a spin quantum number restricted to the single value $s = \frac{1}{2}$:

$$S = \sqrt{s(s+1)}\,\hbar = \frac{\sqrt{3}}{2}\hbar, \qquad S_z = m_s\hbar, \qquad m_s = -\tfrac{1}{2}, +\tfrac{1}{2}. \tag{10.14}$$

The two allowed values of m_s — "spin up" and "spin down" — account exactly for the two Stern–Gerlach spots. A complete specification of an electron's state in hydrogen therefore requires **four** quantum numbers, (n, ℓ, m_ℓ, m_s), and the count of degenerate states for a given n becomes $2n^2$ rather than n^2, the factor of 2 from the two spin orientations — a result central to the structure of the periodic table in Chapter 11.

10.3 Magnetic Moments and Selection Rules

Magnetic Moments

Both orbital and spin angular momentum give the electron a magnetic dipole moment, since a circulating (or intrinsically "spinning") charge behaves as a small current loop. The **orbital magnetic moment** is

$$\vec{\mu}_L = -\frac{e}{2m_e}\vec{L}, \qquad \mu_{L,z} = -m_\ell \mu_B, \tag{10.15}$$

where $\mu_B \equiv e\hbar/2m_e = 9.274 \times 10^{-24}$ J/T is the **Bohr magneton**, a natural unit of atomic magnetic moment. The **spin magnetic moment** has an analogous form but with an extra numerical factor (the electron's g-factor, $g_s \approx 2$, itself a prediction of relativistic quantum theory beyond the scope of the nonrelativistic Schrödinger equation used here):

$$\mu_{S,z} = -g_s\, m_s\, \mu_B \approx -2 m_s \mu_B. \tag{10.16}$$

These magnetic moments are what couple to an external magnetic field in the Stern–Gerlach experiment (producing the observed splitting and deflection) and, coupling to each other and to nuclear magnetic moments, produce the fine and hyperfine splittings observed in high-resolution atomic spectra.

The Stern–Gerlach Force and a Numeric Deflection

A *uniform* magnetic field exerts a torque on a magnetic dipole — producing the Larmor precession pictured in Chapter 9's precessing-top figure — but exerts zero *net force*, because the interaction energy $-\vec{\mu}\cdot\vec{B}$ does not depend on position when $\vec{B}$ itself does not. Deflecting the beam at all, as Stern and Gerlach needed to do to see anything on their screen, requires a spatially varying (inhomogeneous) field, whose gradient converts the dipole's orientation into a genuine transverse force:

$$F_z = \mu_z \frac{\partial B_z}{\partial z}, \tag{10.17}$$

pushing atoms with $\mu_z > 0$ one way and $\mu_z < 0$ the other. Because μ_z takes only the two discrete values $\mu_{S,z} = \pm\mu_B$ (from the spin-moment formula above, with $m_s = \pm\frac{1}{2}$ and $g_s \approx 2$) rather than a continuum of values, the beam splits into exactly two discrete trajectories instead of spreading into one continuous smear — the direct experimental signature of space quantization that motivated this section.

Worked Example: Beam Splitting in a Stern–Gerlach Magnet

A beam of silver atoms ($m = 1.79 \times 10^{-25}$ kg) effuses from an oven at $T = 1000$ K, giving a typical beam speed $v = \sqrt{3k_BT/m} \approx 480$ m/s. The beam passes through a magnet of length $L = 0.50$ m producing a field gradient $\partial B_z/\partial z = 10$ T/m, then drifts a further $D = 0.50$ m to a detector screen. Find the separation between the two spots.

The force on each atom is

$$F_z = \mu_z \frac{\partial B_z}{\partial z} = \pm(9.274 \times 10^{-24}\ \text{J/T})(10\ \text{T/m}) = \pm 9.27 \times 10^{-23}\ \text{N}, \tag{10.18}$$

giving a transverse acceleration $a = F_z/m = \pm 518$ m/s^2. While inside the magnet, for a time $t_1 = L/v = 1.04 \times 10^{-3}$ s, each atom acquires a transverse displacement $z_1 = \frac{1}{2}at_1^2 = 0.28$ mm and a transverse velocity $v_z = at_1 = 0.54$ m/s; it then drifts (no further transverse force, but still moving at v_z) for an additional time $t_2 = D/v = 1.04 \times 10^{-3}$ s, adding $z_2 = v_z t_2 = 0.56$ mm. The one-sided deflection is $z_1 + z_2 \approx 0.84$ mm, so the two spin states land

$$\Delta z = 2(z_1 + z_2) \approx 1.7\ \text{mm} \tag{10.19}$$

apart on the screen — comfortably resolvable, and the same order of magnitude as the splitting Stern and Gerlach actually measured in 1922 with a weaker gradient over a shorter path, which is what made their two-spot pattern (rather than one smeared band) unambiguous.

The Zeeman Effect

Chapter 9 introduced the (normal) **Zeeman effect** qualitatively: an external field $\vec{B}$ along z adds an energy $-\vec{\mu}_L \cdot \vec{B} = m_\ell \mu_B B$ to each otherwise-degenerate m_ℓ sublevel of a given (n, ℓ). Including the spin moment as well, the total field-induced shift of a state (m_ℓ, m_s) is

$$\Delta E = (m_\ell + g_s m_s)\,\mu_B B \approx (m_\ell + 2m_s)\,\mu_B B, \tag{10.20}$$

so that adjacent orbital sublevels are separated by exactly $\mu_B B$. For electric-dipole transitions, the corresponding magnetic quantum-number rule is $\Delta m_\ell = 0, \pm 1$: the $\Delta m_\ell = \pm 1$ transitions produce the two side components, while $\Delta m_\ell = 0$ produces the central component of the normal Zeeman triplet. Historically, it was the *observed* splitting pattern (sometimes the simple three-line "normal" Zeeman pattern predicted by m_ℓ alone, but more often a more complicated "anomalous" pattern explainable only once spin and its distinct g-factor were included) that provided some of the earliest indirect evidence for electron spin, years before Stern and Gerlach identified its direct mechanical signature.

Worked Example: Zeeman Splitting Frequency in a 1 T Field

Estimate the frequency splitting between adjacent m_ℓ sublevels for a hydrogen atom in a $B = 1.0$ T field (a typical laboratory electromagnet), and the resulting wavelength shift of the H_α line ($\lambda = 656.3$ nm, from the worked example above).

$$\Delta E = \mu_B B = (9.274 \times 10^{-24}\ \text{J/T})(1.0\ \text{T}) = 9.27 \times 10^{-24}\ \text{J} = 5.79 \times 10^{-5}\ \text{eV}, \tag{10.21}$$

$$\Delta f = \frac{\Delta E}{h} = \frac{9.27 \times 10^{-24}\ \text{J}}{6.626 \times 10^{-34}\ \text{J}\cdot\text{s}} = 1.40 \times 10^{10}\ \text{Hz} = 14.0\ \text{GHz} \tag{10.22}$$

(the electron's Larmor precession frequency in this field). Converting to a wavelength shift via $|\Delta\lambda| \approx (\lambda^2/c)\,\Delta f$,

$$\Delta\lambda \approx \frac{(656.3 \times 10^{-9}\ \text{m})^2 (1.40 \times 10^{10}\ \text{Hz})}{3.00 \times 10^{8}\ \text{m/s}} \approx 0.020\ \text{nm}, \tag{10.23}$$

about 30,000 times smaller than the line's own wavelength — invisible by eye, but well within reach of a laboratory grating with resolving power $R = \lambda/\Delta\lambda \sim 3 \times 10^4$ (Chapter 5), which is how the Zeeman effect is actually resolved in the lab.

Fine Structure: Spin–Orbit Coupling

Even with no external field at all, the energies $E_n = -13.6\ \text{eV}/n^2$ are not quite the whole story. A useful semiclassical picture is that, in the electron's rest frame, the orbiting (positively charged) nucleus constitutes a circulating current that produces a magnetic field at the electron's location. That internal field couples to the electron's spin magnetic moment, with the necessary relativistic treatment—including Thomas precession—giving the interaction called **spin–orbit coupling**. Its strength depends on the relative orientation of $\vec{L}$ and $\vec{S}$ (equivalently, on the total angular momentum $\vec{J} = \vec{L} + \vec{S}$). Spin–orbit coupling vanishes for $\ell = 0$, but the full **fine structure** also includes relativistic kinetic-energy and Darwin corrections, so s-states can receive fine-structure shifts as well. A complete treatment requires relativistic quantum mechanics beyond the nonrelativistic Schrödinger equation used throughout this book. What matters here is the *size* of the resulting fine-structure correction:

$$\Delta E_{\text{fine}} \sim \alpha^2\,|E_n|, \qquad \alpha \equiv \frac{e^2}{4\pi\epsilon_0\hbar c} \approx \frac{1}{137.0} \approx 7.30 \times 10^{-3}, \tag{10.24}$$

where α, the dimensionless **fine-structure constant**, measures the intrinsic strength of the electromagnetic interaction and sets the size of essentially every relativistic correction in atomic physics. Because $\alpha^2 \approx 5 \times 10^{-5}$, fine structure shifts a level by only a few parts in 10^5 of $|E_n|$ itself — far too fine to see with a simple grating, but readily resolved with a high-resolution spectrometer, and the historical reason spectral lines that look perfectly sharp in an introductory demonstration reveal themselves, under sufficient magnification, to be closely spaced multiplets (hence the name, coined for exactly this kind of fine-toothed splitting of what earlier instruments saw as a single line).

Worked Example: Order-of-Magnitude Fine-Structure Splitting

Estimate the fine-structure splitting of hydrogen's $n = 2$ level, and compare it to the actual measured splitting of the $2p$ levels, $\Delta E_{\text{fine}} \approx 4.5 \times 10^{-5}$ eV (equivalently, about 10.9 GHz).

$$E_2 = -\frac{13.6\ \text{eV}}{4} = -3.40\,\text{eV} \quad\Longrightarrow\quad \Delta E_{\text{fine}} \sim \alpha^2 |E_2| = (7.30{\times}10^{-3})^2 (3.40\,\text{eV}) \approx 1.8{\times}10^{-4}\,\text{eV}, \tag{10.25}$$

within about a factor of four of the measured value — entirely appropriate for an order-of-magnitude estimate that drops the numerical factors (j-dependence and other combinatorics supplied only by a full relativistic treatment) while correctly capturing the essential physical scale: fine structure is smaller than the gross level spacing $|E_n|$ by a factor of $\alpha^2 \approx 5 \times 10^{-5}$, the same small parameter that controls, order by order, essentially every relativistic correction in atomic physics.

The same coupling, applied to the *proton's* magnetic moment rather than the electron's, is the basis of magnetic resonance imaging, and Figure 10.5 is a working model of it. A static field along z splits the two spin orientations by $\Delta E = g\mu_N B$, which for a proton in a 1 T field falls in the radio band; a transverse radio-frequency field tuned to exactly that frequency drives transitions between them, and the resonance is sharp enough that a deliberate spatial *gradient* in B makes the resonant frequency a map of position. Everything in the sequence — the splitting proportional to B, the resonance condition $hf = \Delta E$, the return to equilibrium afterwards — is this section's physics with μ_B replaced by the nuclear magneton.

Nuclear magneton. $\mu_N \equiv e\hbar/2m_p$, built the same way as the Bohr magneton μ_B but with the much larger proton mass m_p in place of m_e. Since $m_p \approx 1836\, m_e$, $\mu_N \approx \mu_B/1836$ — nuclear magnetic effects are intrinsically about a thousand times weaker than electronic ones, which is part of why MRI needs strong fields and sensitive receivers.

Selection Rules

Not every pair of hydrogen energy levels is connected by an observable spectral line. An electron making a transition between stationary states typically does so by emitting or absorbing a single photon, and conservation of angular momentum and parity restricts which transitions can occur via this single-photon (electric dipole) process. The principal orbital **selection rule** is

$$\Delta\ell = \pm 1, \tag{10.26}$$

and the magnetic quantum number must obey $\Delta m_\ell = 0, \pm 1$; the electron's spin projection is unchanged, $\Delta m_s = 0$. There is no similarly strict restriction on Δn, but the electric-dipole transition must change parity, which the $\Delta\ell = \pm 1$ rule already guarantees for hydrogen orbitals. Transitions violating these rules (e.g., $2s \to 1s$, both $\ell = 0$) are called **forbidden transitions** — not absolutely impossible, but strongly suppressed, occurring (if at all) only through much slower, higher-order processes. The selection rules are why, for instance, the observed hydrogen spectral series (Lyman, Balmer, Paschen, etc., corresponding to transitions ending on $n_f = 1, 2, 3, \ldots$) show specific line patterns rather than a line for every conceivable pair of levels.

Tip

When checking whether a transition is allowed, check $\Delta\ell$ first and don't spend time on Δn — the principal selection rule places no restriction on n. Then check Δm_ℓ if the magnetic sublevels are specified. A quick way to keep this straight while scanning a level diagram grouped by orbital letter: an allowed electric-dipole transition always moves exactly one letter over ($s \leftrightarrow p$, $p \leftrightarrow d$, etc.), never staying within the same letter and never skipping one, regardless of how far it jumps between n values.

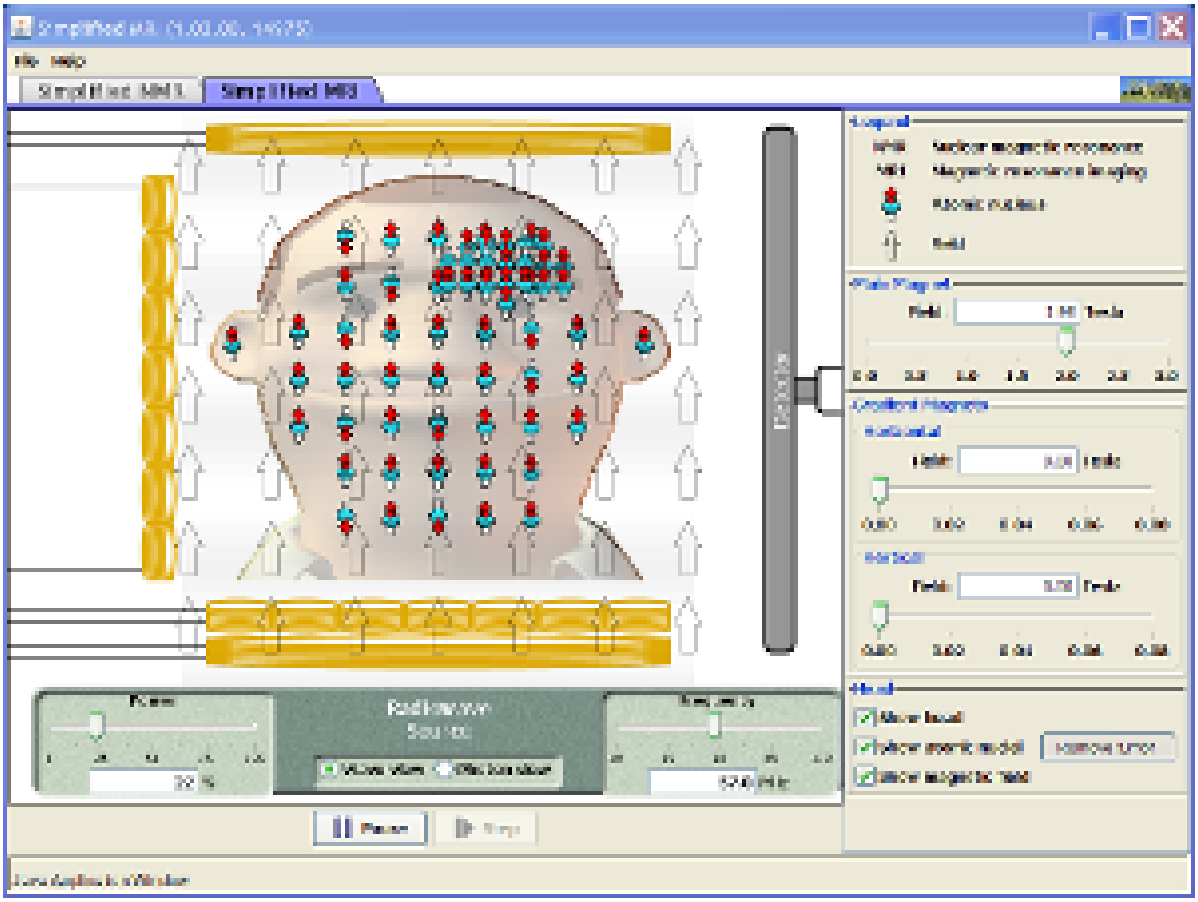

Figure 10.5: Proton spins in a magnetic field: Zeeman splitting, resonant absorption of a radio-frequency photon, and the field gradient that turns the resonance into an image. Interactive simulation: Simplified MRI

10.4 Summary

- The radial Schrödinger equation for hydrogen's $1/r$ Coulomb potential reproduces the Bohr energy levels $E_n = -13.6\ \text{eV}/n^2$, but from normalizability of $R(r)$ rather than a postulated orbit, and with the ground state having zero orbital angular momentum, unlike Bohr's model. Normalizability forces the radial series to terminate into a finite polynomial, which fixes n to a positive integer and, via $n = n_r + \ell + 1$ ($n_r \geq 0$ the radial node count), caps ℓ at $n - 1$.
- **Bohr's 1913 model** reproduced hydrogen's spectrum numerically from an unjustified orbit postulate, but could not be extended consistently to helium or predict transition rates, and gets the ground state's angular momentum flatly wrong — failures that motivated the full Schrödinger treatment.
- **Hydrogenic ions** (e.g. He^+, Li^{2+}) follow $E_n(Z) = -Z^2(13.6\ \text{eV})/n^2$ and $a_n(Z) = n^2 a_0/Z$: binding energy scales as Z^2, orbital size as $1/Z$.
- Hydrogen states are labeled by n, $\ell = 0, \ldots, n-1$, $m_\ell = -\ell, \ldots, \ell$; because E_n depends only on n, there are n^2 degenerate spatial states per n — a special feature of the pure Coulomb potential.
- Explicit radial wave functions R_{10}, R_{20}, R_{21}, and the radial probability distributions $P(r) = r^2|R_{n\ell}|^2$ they generate, show the $n - \ell - 1$ node count directly; angular nodal structure gives rise to the characteristic s, p, d orbital shapes. The Bohr radius a_0 reappears as the most probable electron-nucleus separation in the ground state.
- The **effective potential** $V_{\text{eff}}(r) = V(r) + \hbar^2\ell(\ell+1)/2m_e r^2$ shows that only $\ell = 0$ states have nonzero probability density at the nucleus (although the radial probability $P(r)$ vanishes there for every state); higher ℓ states are held away from the origin by the centrifugal barrier, a penetration effect central to screening in Chapter 11.
- The **Lyman**, **Balmer**, and **Paschen** series (transitions ending on $n_f = 1, 2, 3$) each converge to a series limit at $n_i \to \infty$; the Rydberg formula $1/\lambda = R(1/n_f^2 - 1/n_i^2)$ gives their wavelengths, with the Balmer series (364.6–656.3 nm) spanning most of the visible range.
- The **Stern–Gerlach experiment** revealed **electron spin**, an intrinsic angular momentum with quantum number $s = \frac{1}{2}$ and $m_s = \pm\frac{1}{2}$, doubling the degenerate state count to $2n^2$ and requiring four quantum numbers (n, ℓ, m_ℓ, m_s) to fully specify a state. The inhomogeneous-field force $F_z = \mu_z\, \partial B_z/\partial z$ converts the two spin orientations into two discrete beam deflections.
- Orbital and spin angular momentum each produce a magnetic moment, in units of the Bohr magneton μ_B; the spin moment carries an extra g-factor ≈ 2. In an external field, the (normal) **Zeeman effect** shifts sublevels by $\Delta E \approx (m_\ell + 2m_s)\mu_B B$; even with no external field, **spin–orbit coupling** produces **fine-structure** splitting of order $\alpha^2|E_n|$, where $\alpha \approx 1/137$ is the fine-structure constant.
- Single-photon electric-dipole transitions obey $\Delta\ell = \pm 1$, $\Delta m_\ell = 0, \pm 1$, and $\Delta m_s = 0$, explaining the observed pattern of hydrogen spectral lines.

10.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens on the hydrogen energy-level diagram.

1. **Multiple choice, from a figure.** Which orbital selection rule governs an ordinary electric-dipole transition? Figure description: Hydrogen energy levels and the Lyman, Balmer, and Paschen transition series. Choices: (A) $\Delta\ell = \pm 1$. (B) $\Delta\ell = 0$ only. (C) $\Delta n = 0$ only. (D) $\Delta ms = \pm 1$ is required.

2. **True or false.** The hydrogen ground state has orbital angular momentum $\hbar$, as in the original Bohr model.

3. **Drag the words.** Complete the hydrogen quantum-number rules. For a given n, ℓ runs from _____. For a given ℓ, $m\ell$ runs from _____. The spin projection is _____. Electric-dipole transitions require _____. (Terms: 0 to n - 1; $-\ell$ to $+\ell$; ms = $\pm 1/2$; $\Delta\ell = \pm 1$; 1 to n; $\Delta\ell = 0$.)

4. **Fill in the blanks.** Complete the hydrogenic scaling at fixed n. Increasing nuclear charge from Z = 1 to Z = 2 makes the binding-energy magnitude _____ times as large and the characteristic orbital radius _____ as large.

5. **Mark the words.** Mark the two results revealed by the Stern–Gerlach experiment. "The beam separates into two discrete deflections, revealing an intrinsic spin-one-half degree of freedom. It does not produce a continuous range of orientations."

10.6 Conceptual Questions

1. Why does a hydrogen atom have discrete bound energies but a continuum above ionization?

2. Distinguish radial nodes from angular nodes and relate each count to the quantum numbers.

3. Why do s orbitals penetrate the nuclear region more strongly than orbitals with $\ell > 0$?

4. Explain why the Coulomb spectrum is degenerate in ℓ in the nonrelativistic hydrogen model.

5. Why did the Stern–Gerlach result require a new two-valued angular momentum rather than ordinary orbital angular momentum alone?

6. A transition can conserve energy but still be forbidden by a selection rule. Explain what "forbidden" means experimentally.

10.7 Problems

Exercise 10.1
(Straightforward)

Using the Rydberg formula, compute the wavelength of the first line of the Lyman series ($n = 2 \to n = 1$) and the Lyman series limit ($n \to \infty \to n = 1$). State which portion of the electromagnetic spectrum both lie in, and explain why the entire Lyman series (unlike the Balmer series) is invisible to the human eye.

Exercise 10.2
(Straightforward)

List all allowed (ℓ, m_ℓ) combinations for $n = 3$, count the total number of spatial states, and verify this equals $n^2 = 9$. Including spin, how many total quantum states share this energy?

Exercise 10.3
(Straightforward)

Using the selection rule $\Delta\ell = \pm 1$, determine which of the following single-photon transitions are allowed and which are forbidden: (a) $3d \to 2p$, (b) $3s \to 2s$, (c) $3p \to 1s$, (d) $2p \to 1s$.

Exercise 10.4
(Straightforward)

An electron in a hydrogen atom is in a $3d$ state ($\ell = 2$). (a) Compute the magnitude of its orbital angular momentum in units of $\hbar$. (b) Compute the largest possible magnitude of its z-component orbital magnetic moment, in units of the Bohr magneton.

Exercise 10.5
(Moderate)

In the Stern–Gerlach experiment, explain why silver atoms (rather than, say, helium atoms) were a good choice for demonstrating space quantization due to electron spin, referring to the electron configuration of the outermost electron (you may look ahead to Chapter 11's discussion of electron configurations, or simply reason from the fact that silver's single outer electron is in an s-state).

Exercise 10.6
(Moderate)

Show that the ground-state radial probability distribution of hydrogen, $P(r) = r^2|R_{10}(r)|^2 \propto r^2 e^{-2r/a_0}$, is maximized at $r = a_0$ by differentiating $P(r)$ with respect to r and setting the result to zero.

Exercise 10.7
(Moderate)

Doubly ionized lithium, Li^{2+} ($Z = 3$), is a one-electron hydrogenic ion. (a) Find its ground-state ionization energy. (b) Find the radius at which its ground-state radial probability distribution peaks. (c) Compare both results, as ratios, to hydrogen's, and confirm they follow the Z^2 and $1/Z$ scaling laws given in the text.

Exercise 10.8
(Moderate)

A different Stern–Gerlach apparatus uses a field gradient $\partial B_z/\partial z = 15$ T/m over a magnet of length $L = 0.20$ m, with silver atoms ($m = 1.79 \times 10^{-25}$ kg) moving at $v = 600$ m/s, followed by a $D = 0.40$ m drift to the screen. Using the method of the worked example, find the separation between the two spots.

Exercise 10.9

(Challenging)

A hydrogen discharge tube is placed in a magnetic field of $B = 0.50$ T. (a) Find the Zeeman energy splitting ΔE (in eV) between adjacent m_ℓ sublevels. (b) Find the corresponding frequency splitting Δf. (c) Using $\lambda = 486.1$ nm (the H_β line), estimate the wavelength splitting $\Delta\lambda$, and compare its order of magnitude to the 0.50 T case's expected scaling relative to the 1.0 T worked example in the text.

Exercise 10.10

(Challenging)

Estimate the order-of-magnitude fine-structure splitting of hydrogen's $n = 3$ level, using $\Delta E_{\text{fine}} \sim \alpha^2 |E_n|$. Compare your result to the $n = 2$ estimate found in the text, and explain, in terms of the $|E_n| = 13.6\ \text{eV}/n^2$ scaling, why fine structure becomes progressively harder to resolve at higher n.

Chapter 11

Many-Electron Atoms

Learning Objectives

By the end of this chapter, you should be able to:

- State the Pauli exclusion principle and explain its role in determining atomic structure.
- Explain why the energy of a state in a multi-electron atom depends on both n and ℓ, unlike in hydrogen, using the concept of electron screening.
- Write the ground-state electron configuration of a given element using the subshell-filling order and Hund's rule.
- Relate the structure of the periodic table (periods, groups, blocks) to electron configurations.
- Explain the origin of characteristic X-ray spectra and apply Moseley's law.
- Describe the physical basis of the laser: stimulated emission, population inversion, and metastable states.

Introduction

Chapter 10 solved the hydrogen atom exactly, obtaining energies that depend on a single quantum number n and states that can be filled with electrons two at a time (spin up, spin down) up to a degeneracy of $2n^2$. This chapter asks what happens for atoms with more than one electron, where the Schrödinger equation cannot be solved exactly, because each electron interacts not only with the nucleus but with every other electron. Two ideas make the many-electron problem tractable and explain the entire structure of the periodic table: an *independent-particle approximation*, in which each electron is imagined to move in an effective, average potential created by the nucleus and all the other electrons, and the *Pauli exclusion principle*, a rule with no classical analog that limits how many electrons can occupy any single quantum state. Together they explain why elements have the chemical and spectroscopic properties they do — and, in doing so, resolve one of the oldest puzzles in physics that quantum mechanics was built to explain: the periodic table itself.

11.1 The Exclusion Principle and Screening

The Pauli Exclusion Principle

Wolfgang Pauli proposed in 1925 (before the discovery of the Schrödinger equation) a rule required to explain observed atomic spectra and, later, understood as a consequence of the

fundamentally indistinguishable and antisymmetric nature of electron wave functions: **no two electrons in an atom can occupy the same complete set of quantum numbers** (n, ℓ, m_ℓ, m_s). Equivalently, each distinct spatial-and-spin quantum state (n, ℓ, m_ℓ, m_s) can hold **at most one electron**. This is not a subtle statistical tendency but an absolute prohibition, and it applies generally to electrons (and more broadly to the class of particles called fermions, which includes protons and neutrons) — without it, every electron in a multi-electron atom could simply fall into the lowest-energy $1s$ state, and all atoms would have similar, small sizes and similar chemistry, in sharp contradiction to the observed diversity of the periodic table. The exclusion principle is the single most important input, beyond the Schrödinger equation itself, needed to explain atomic structure.

Pauli's Nobel Prize — and the Deeper Theorem Behind the Rule

Pauli received the 1945 Nobel Prize in Physics "for the discovery of the Exclusion Principle," on a nomination by Albert Einstein — two decades after the 1925 postulate itself, and a reminder of how long it can take for a rule with no classical mechanism behind it to be fully vindicated. Pauli did not stop at the empirical rule: in 1940 he proved the **spin-statistics theorem**, showing that any relativistic quantum field theory *requires* half-integer-spin particles (fermions, including electrons) to obey antisymmetric statistics and integer-spin particles (bosons, including photons) to obey symmetric statistics. The exclusion principle, in other words, is not a free-standing postulate about electrons specifically — it is a necessary consequence of relativity and quantum field theory applied to any spin-$\frac{1}{2}$ particle.

Historical Context: Pauli's Reasoning and Hund's Empirical Rule

Pauli arrived at the exclusion principle in late 1924 by a route that had nothing to do with wave mechanics, which did not yet exist — Schrödinger's equation was still two years away. Working instead from the older Bohr–Sommerfeld quantum theory, Pauli was trying to explain two stubborn regularities that orbital quantum numbers alone could not account for: the precise electron counts (2, 8, 8, 18, ...) at which successive shells close to give a noble gas, and the "anomalous" Zeeman splitting of certain spectral lines into a number of components inconsistent with any known combination of n and ℓ. Pauli proposed that the electron carries an additional, two-valued property — which he described, before its physical meaning as spin was proposed by Goudsmit and Uhlenbeck the following year, only as a "classically non-describable two-valuedness" — and that no two electrons in an atom may share an identical set of all their quantum numbers, orbital and this new one included. Both puzzles fell into place at once: the shell-closing numbers became simple counting exercises, and the anomalous splittings became ordinary Zeeman splittings of a spin degree of freedom nobody had previously included.

Hund's rule has a similarly empirical origin. Friedrich Hund formulated it around 1925 not from any first-principles calculation but by systematically cataloging the term symbols (the quantum numbers characterizing the total orbital and spin angular momentum) that best fit the observed ground-state spectra of many atoms, and noticing that the ground state was consistently the one with maximum total spin. Only later, once the Schrödinger equation and the antisymmetry of many-electron wave functions were understood, did it become clear *why*: electrons with parallel spins are already kept apart by the exclusion principle itself (their spatial wave function must be antisymmetric, vanishing whenever two such electrons approach the same point), which lowers their mutual Coulomb repulsion energy compared to electrons of opposite spin occupying the same orbital. This spin-correlated reduction in repulsion is called

the **exchange energy**, and it reappears below as the reason certain elements (chromium and copper among them) deviate from the naive Aufbau filling order.

A **term symbol**, written ${}^{2S+1}L_J$, packages an atomic state's total spin S, total orbital angular momentum L (coded $S, P, D, F, \ldots$ for $L = 0, 1, 2, 3, \ldots$), and total angular momentum J into one label; Hund caught the *ground-state* pattern of maximum S among these labels well before anyone could compute them from first principles.

The Real Content of the Exclusion Principle: Antisymmetric Wave Functions

Stated as "no two electrons can share the same quantum numbers," the exclusion principle can look like an arbitrary bookkeeping rule bolted onto quantum mechanics. Its actual origin is a deeper and more general statement about *identical particles*: because electrons are fundamentally indistinguishable, swapping the labels of any two electrons in a multi-electron wave function must leave every measurable prediction unchanged, which requires the total wave function to be either symmetric or antisymmetric under that exchange. Experiment (and the spin-statistics theorem) fixes electrons, as spin-$\frac{1}{2}$ fermions, to the antisymmetric choice.

For two electrons in single-particle states ψ_a and ψ_b, the properly antisymmetrized two-electron wave function is

$$\Psi(1,2) = \frac{1}{\sqrt{2}}\Big[\psi_a(1)\psi_b(2) - \psi_a(2)\psi_b(1)\Big]. \tag{11.1}$$

Swapping the labels $1 \leftrightarrow 2$ multiplies Ψ by -1, as required of a fermion wave function. Now suppose the two states are the same, $a = b$: the expression becomes $\frac{1}{\sqrt{2}}[\psi_a(1)\psi_a(2) - \psi_a(2)\psi_a(1)] = 0$ identically. The wave function for two electrons in the same quantum state does not merely describe an unlikely configuration — it vanishes outright, for every possible position and spin of both electrons. "No two electrons occupy the same state" is not a separate postulate; it is this algebraic fact, stated in words. The same antisymmetry, applied to two electrons in *different* orbitals with parallel spin, is also what produces the exchange energy discussed above: the antisymmetric spatial wave function is forced toward zero whenever the two electrons approach the same point, keeping same-spin electrons farther apart on average and lowering their mutual Coulomb repulsion.

Screening and Subshell Energies

In hydrogen, the energy of a state depends only on n (Chapter 10), because the electron feels the bare $1/r$ potential of a single proton. In a multi-electron atom, an electron in an outer shell is partially **screened** from the full nuclear charge Ze by the electrons in shells closer to the nucleus: it feels an *effective* nuclear charge $Z_{\text{eff}}e < Ze$, reduced from the true charge by the (partial) shielding effect of the intervening electron cloud.

A **shell** groups states by principal quantum number n alone; a **subshell** further groups by ℓ within a shell (e.g., $n = 3$ is a shell; $3s$, $3p$, $3d$ are its three subshells). Screening is what makes subshell, not just shell, energy physically meaningful.

Screening depends on ℓ as well as n, because electrons of lower ℓ (at fixed n) have wave functions with a greater probability of being found close to the nucleus (their radial probability distributions extend closer to $r = 0$, as can be seen in the general shape of the hydrogen radial

functions of Chapter 10) — such electrons penetrate the inner electron cloud more effectively, feel less screening, and are therefore more tightly bound. The result is that, unlike in hydrogen, **energy in a multi-electron atom depends on both n and ℓ**, with energy generally increasing with ℓ at fixed n: within a given n, an s state ($\ell = 0$) lies lower in energy than a p state ($\ell = 1$), which lies lower than a d state ($\ell = 2$), and so on. This ℓ-dependence is what breaks hydrogen's accidental degeneracy and is responsible for the specific subshell-filling order used below.

The Self-Consistent Field: How Screening Is Actually Computed

The qualitative picture above — "an electron feels a reduced effective charge $Z_{\text{eff}}e$" — can be made into an actual calculation, and doing so is instructive even though the resulting numbers are only ever approximate. Douglas Hartree, in 1928, proposed treating each electron as moving independently in an *effective, spherically averaged* potential built from the nucleus plus the smeared-out charge distribution of every other electron: $V_{\text{eff}}(r) = -\dfrac{Ze^2}{4\pi\epsilon_0 r} + V_{\text{other electrons}}(r)$. The trouble is circular — computing $V_{\text{other electrons}}(r)$ requires already knowing the wave functions of all the other electrons, which is exactly what one is trying to find. Hartree's resolution, the **self-consistent field (SCF) method**, is an iterative procedure: start with a reasonable guess for every electron's wave function (hydrogen-like orbitals, say); use those guesses to compute an averaged charge density and hence an effective potential $V_{\text{eff}}(r)$ for each electron; solve the resulting one-electron Schrödinger equation for each electron in that potential, obtaining *improved* wave functions; recompute the charge density and the effective potential from these improved wave functions; and repeat. The cycle is stopped once the potential fed in and the potential computed out agree to the desired precision — the field is then "self-consistent," and the resulting energies and wave functions are the Hartree approximation's best estimate for the atom's structure. (A refinement due to Vladimir Fock in 1930, Hartree–Fock theory, additionally enforces the antisymmetry required by the exclusion principle itself, which plain Hartree theory omits, and is the starting point for essentially all modern atomic-structure and quantum-chemistry calculations.)

This procedure makes the r-dependence of screening explicit rather than assumed: an electron that penetrates *inside* the charge cloud of the other electrons (small r) sees a potential close to the full, unscreened nuclear charge Z, while an electron that stays *outside* essentially all the other electrons (large r) sees a much-reduced net charge — roughly Z minus the number of electrons enclosed within its orbit. A full SCF calculation is impractical by hand, which is why **Slater's rules** (1930) exist: an empirical, hand-computable recipe, calibrated against Hartree-type results, for estimating a single number Z_{eff} for a given electron without running the iteration at all.

Worked Example: Effective Nuclear Charge via Slater's Rules

Slater's rules group orbitals as $(1s)$, $(2s, 2p)$, $(3s, 3p)$, $(3d)$, $(4s, 4p)$, … and estimate the total screening S felt by an electron in an ns or np orbital as a sum of contributions: 0.35 for each *other* electron in the same group (0.30 if the group is $1s$), 0.85 for each electron one shell lower ($n-1$), and 1.00 (full screening) for each electron two or more shells lower. Then $Z_{\text{eff}} = Z - S$.

For the single valence electron of **sodium** ($Z = 11$, configuration $1s^2 2s^2 2p^6 3s^1$), the electron of interest is the lone $3s$ electron. There are no other electrons in its $(3s, 3p)$ group; the $n = 2$ shell contributes 8 electrons ($2s^2 2p^6$) at 0.85 each; the $n = 1$ shell contributes 2 electrons at 1.00 each:

$$S = 8(0.85) + 2(1.00) = 6.80 + 2.00 = 8.80, \qquad Z_{\text{eff}} = 11 - 8.80 = 2.20. \tag{11.2}$$

For a $3p$ electron of **chlorine** ($Z = 17$, configuration $1s^2 2s^2 2p^6 3s^2 3p^5$), the same-group electrons are the remaining 6 electrons of $3s^2 3p^5$ (excluding the one of interest), the $n = 2$ shell again contributes 8 electrons at 0.85, and the $n = 1$ shell contributes 2 at 1.00:

$$S = 6(0.35)+8(0.85)+2(1.00) = 2.10+6.80+2.00 = 10.90, \qquad Z_{\text{eff}} = 17-10.90 = 6.10. \tag{11.3}$$

Moving across period 3 from sodium to chlorine, the true nuclear charge grows by 6 protons, but the added electrons (all in the *same* shell) screen each other only weakly (0.35 each rather than 0.85 or 1.00), so Z_{eff} very nearly *keeps pace* with Z, rising from 2.20 to 6.10 — nearly a threefold increase. This is the quantitative content behind the qualitative statement that effective nuclear charge rises sharply across a period, and it is the mechanism explored further below in connection with ionization energy and atomic size.

11.2 Electron Configurations and the Periodic Table

Combining the exclusion principle with the n,ℓ-dependent ordering of subshell energies, the ground-state **electron configuration** of an atom is built by filling the lowest-energy available subshells first, two electrons (spin up and spin down) per orbital, up to $2(2\ell + 1)$ electrons per subshell — this is the **Aufbau ("building-up") principle**. Because screening shifts subshell energies, the filling order does not simply follow increasing n; the empirical (and largely first-principles-derivable) order is approximately

$$1s,\ 2s,\ 2p,\ 3s,\ 3p,\ 4s,\ 3d,\ 4p,\ 5s,\ 4d,\ 5p,\ 6s,\ 4f,\ 5d,\ 6p,\ \ldots \tag{11.4}$$

— note, for example, that $4s$ fills before $3d$, since the extra penetration of the $4s$ orbital lowers its energy below that of $3d$ despite its larger n. When a subshell contains more than one electron and is only partially filled, **Hund's rule** states that the ground-state configuration maximizes the total spin (electrons singly occupy separate orbitals within a subshell, with parallel spins, before any orbital is doubly occupied) — a consequence of electron-electron repulsion, which is minimized when electrons, to the extent the exclusion principle allows, avoid occupying the same spatial orbital.

> **Core** electrons occupy filled inner subshells and are usually not involved in ordinary bonding; **valence** electrons are the outermost electrons and, for transition or inner-transition atoms, can also include nearby incompletely filled d or f subshells. These are the electrons most directly involved in bonding and reactivity.

This filling scheme directly generates the structure of the **periodic table**. Each **period** (row) begins filling a new principal shell n; in the transition and inner-transition blocks, however, a lower-n d or f subshell is filled after the outer ns subshell, so a period is not simply a single-shell filling sequence. Each period ends when every subshell filled so far is completely full, the particularly stable arrangement that defines a noble gas. Elements in the same **group** (column) generally share the same outer-shell (**valence**) configuration and, correspondingly, similar chemical properties, since chemical bonding (Chapter 12) is governed primarily by the valence electrons. The table's division into s-block, p-block, d-block (transition metals), and

f-block (lanthanides/actinides) regions directly reflects which subshell is being filled across that block. The chemical inertness of the noble gases, the strong reactivity of the alkali metals (a single, loosely bound s-electron outside a filled shell) and the halogens (one electron short of a filled shell), and the broad periodicity of atomic size and ionization energy all follow from this shell structure, without further assumptions.

Pauli Exclusion Beyond Atoms: The Nuclear Shell Model

The exclusion principle is not a special rule about electrons in atoms — it applies to any collection of identical fermions. Protons and neutrons are themselves spin-$\frac{1}{2}$ fermions, and each species independently fills its own sequence of energy levels inside the nucleus subject to the same exclusion principle, producing a shell structure with its own closed-shell "magic numbers" that is strikingly analogous to the noble-gas shell closures described here. See Nuclear Physics for the nuclear shell model built on exactly this idea.

The Aufbau procedure is quicker to learn by doing it than by reading the rules, and Figure 11.1 is the exercise: add protons, neutrons, and electrons one at a time and watch the element name, the net charge, the mass number, and the shell occupancy update together. Two habits of thought are worth breaking there. Adding a proton changes the element; adding a neutron does not, and only moves the atom along a row of isotopes (Chapter 13). And the electrons go into shells that fill in a fixed order, so that the chemistry of the atom is decided by the last few added rather than by the total.

The Aufbau Order in Practice: Configurations Across the Periodic Table

The filling order $1s, 2s, 2p, 3s, 3p, 4s, 3d, 4p, \ldots$ is best seen applied across a representative spread of the periodic table, rather than for one or two elements in isolation. Table 11.1 builds ground-state configurations by strict Aufbau filling for most entries, but flags two elements — **chromium** and **copper** — where the real, spectroscopically measured ground state disagrees with the naive prediction, and includes **lanthanum**, the first element for which the $4f$ subshell becomes energetically relevant, as a further illustration of just how close some subshell energies run to one another.

Two features stand out. First, every noble gas (He, Ne, Ar, Kr) ends a row with every subshell up to that point completely filled — the defining structural feature of a noble gas, and the reason the next electron (starting a new period) must go into a new, much less tightly bound shell, exactly the drop in binding responsible for the alkali metals' low ionization energy discussed below. Second, three of the eleven entries in the table are Aufbau *exceptions*, which is a useful reminder that the filling order given earlier in this chapter is a good approximation, not an exact law — the true ordering is whatever a full self-consistent-field calculation (or, in practice, the measured spectrum) says it is, and the mnemonic order is simply the pattern that calculation follows *most* of the time.

Spotting Aufbau Exceptions Before You Look Them Up

The naive Aufbau order gets the ground-state configuration right for the overwhelming majority of elements, so it is worth trusting by default — but it is worth pausing to check whenever straightforward filling would leave a d or f subshell *one electron short* of exactly half-filled (d^4, f^6) or exactly one electron short of completely filled (d^9, f^{13}). Those are

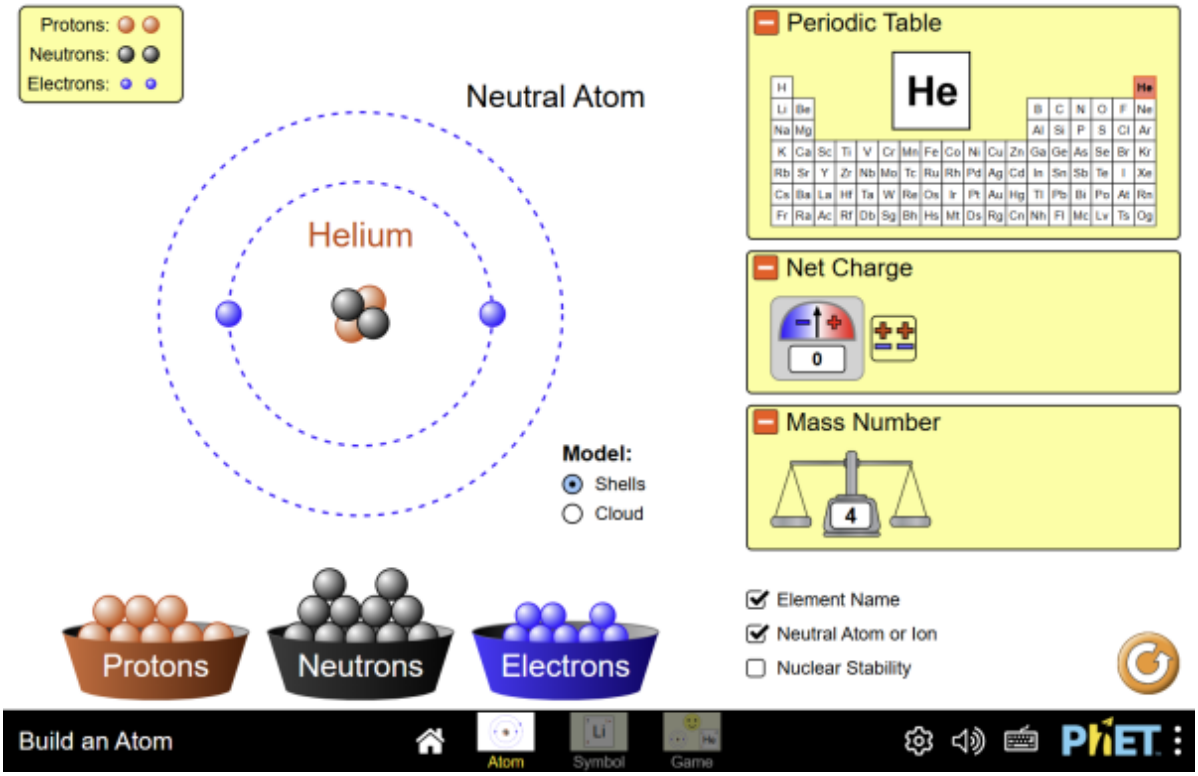

Figure 11.1: Atoms assembled particle by particle, with element, charge, mass number, and electron shell filling all displayed as they are built. Build up to neon or argon and the outer shell closes exactly as the electron count reaches a noble gas — the closure being a fact about the electrons, and nothing to do with the nucleus underneath them. Interactive simulation: Build an Atom

Table 11.1: Ground-state electron configurations across the periodic table, by strict Aufbau filling order.

Element	Z	Naive Aufbau prediction	Actual ground state
Helium	2	$1s^2$	$1s^2$
Neon	10	$1s^2 2s^2 2p^6$	$1s^2 2s^2 2p^6$
Argon	18	$[\mathrm{Ne}]\, 3s^2 3p^6$	$[\mathrm{Ne}]\, 3s^2 3p^6$
Potassium	19	$[\mathrm{Ar}]\, 4s^1$	$[\mathrm{Ar}]\, 4s^1$
Calcium	20	$[\mathrm{Ar}]\, 4s^2$	$[\mathrm{Ar}]\, 4s^2$
Scandium	21	$[\mathrm{Ar}]\, 3d^1 4s^2$	$[\mathrm{Ar}]\, 3d^1 4s^2$
Chromium	24	$[\mathrm{Ar}]\, 3d^4 4s^2$	$[\mathrm{Ar}]\, 3d^5 4s^1$ (exception)
Iron	26	$[\mathrm{Ar}]\, 3d^6 4s^2$	$[\mathrm{Ar}]\, 3d^6 4s^2$
Copper	29	$[\mathrm{Ar}]\, 3d^9 4s^2$	$[\mathrm{Ar}]\, 3d^{10} 4s^1$ (exception)
Krypton	36	$[\mathrm{Ar}]\, 3d^{10} 4s^2 4p^6$	$[\mathrm{Ar}]\, 3d^{10} 4s^2 4p^6$
Lanthanum	57	$[\mathrm{Xe}]\, 4f^1 6s^2$	$[\mathrm{Xe}]\, 5d^1 6s^2$ (exception)

precisely the configurations where promoting one electron from the outer s subshell buys extra exchange-energy stabilization by reaching d^5, d^{10}, f^7, or f^{14}. If your predicted configuration lands one electron short of one of those special counts, it is worth double-checking against a reference table before reporting it as the ground state.

Worked Example: Why Chromium's Ground State Breaks the Naive Aufbau Order

The Aufbau order predicts chromium ($Z = 24$) should have configuration $[\mathrm{Ar}]\,3d^4 4s^2$: fill $4s$ completely (it is lower in energy than $3d$, as established above), then place the remaining four electrons in $3d$. The measured ground state is instead $[\mathrm{Ar}]\,3d^5 4s^1$ — one electron has been promoted from $4s$ down into $3d$, apparently *uphill* in single-particle energy.

The resolution is that single-particle subshell energies are not the whole story; the *total* energy of the atom also includes the electron-electron interaction energy, and that interaction is lower, for a fixed set of occupied orbitals, when as many electrons as possible have parallel spin (the exchange energy discussed in the Historical Context box above). The configuration $3d^5 4s^1$ places one electron in *each* of the five $3d$ orbitals ($m_\ell = -2, -1, 0, 1, 2$) with parallel spin, and the $4s$ electron can align with them as well. In the naive $3d^4 4s^2$ configuration, only four $3d$ orbitals are singly occupied, while the $4s$ orbital is paired and the fifth $3d$ orbital is empty. The half-filled $3d$ subshell therefore gains substantial exchange stabilization, and for chromium specifically that extra lowering outweighs the modest single-particle energy cost of moving one electron from $4s$ up into $3d$. Copper ($Z = 29$) is analogous: $[\mathrm{Ar}]\,3d^{10} 4s^1$ trades a doubly occupied $4s$ for a completely filled, maximally stable $3d^{10}$ shell (all orbitals doubly occupied — no further exchange gain from that shell alone, but a filled subshell carries its own extra stability, being spherically symmetric and having zero net orbital angular momentum) rather than the naive $3d^9 4s^2$. These familiar exceptions occur at *half-filled* (d^5) and *completely filled* (d^{10}) subshells, where exchange and subshell-closure effects can favor the promotion.

Exchange Energy Is Not a Classical Force

It is tempting to picture the exchange-energy stabilization behind Hund's rule and the chromium/copper exceptions as electrons with parallel spin somehow "repelling less" because of some spin-dependent force between them, the way two bar magnets attract or repel depending on orientation. That picture is wrong: the electron-electron interaction here is the ordinary electrostatic Coulomb repulsion, and there is no additional spin-based force anywhere in the Hamiltonian. The energy difference is a purely statistical consequence of antisymmetry — the antisymmetric *spatial* wave function required for two parallel-spin electrons is forced toward zero as the two electrons approach each other, so they simply spend less time close together, and therefore have lower average Coulomb repulsion, than two opposite-spin electrons whose (symmetric) spatial wave function carries no such restriction. Same-spin electrons are not pushed apart by a force; they are correlated apart by the requirement of exchange antisymmetry.

Ionization Energy and Atomic Radius Across the Periodic Table

The Slater's-rule calculation above — Z_{eff} rising from 2.20 (sodium) to 6.10 (chlorine) across period 3 — is the quantitative content behind two of the most familiar patterns in chemistry.

Atomic radius decreases across a period: each added electron goes into the *same* shell (same n, hence similar average distance from the nucleus, all else equal), while Z_{eff} pulling on that shell steadily increases, drawing the whole electron cloud inward. **First ionization energy** — the energy required to remove the least tightly bound electron — correspondingly *increases* across a period, since a more tightly bound (higher Z_{eff}, smaller-radius) electron is harder to remove; sodium's outermost electron ($I_1 = 5.14$ eV) is far easier to strip away than chlorine's ($I_1 = 12.97$ eV), consistent with the roughly threefold increase in Z_{eff} found above. Both trends reverse on descending a group: each new period adds a shell of larger n, and although Z_{eff} for the outermost electron changes only mildly down a group (the added inner shells screen the added protons quite effectively, shell for shell), the larger n dominates, so atomic radius grows and ionization energy falls — lithium, sodium, and potassium have successively *lower* first ionization energies (5.39, 5.14, 4.34 eV) even though Z more than sextuples between them.

Figure 11.2 shows the resulting pattern for every element from hydrogen through krypton: not a smooth trend but a sharp periodic sawtooth, rising across each period and collapsing at the start of the next, with a noble gas at every peak and an alkali metal at every trough — the periodic table's defining periodicity made directly visible in a single measurable quantity.

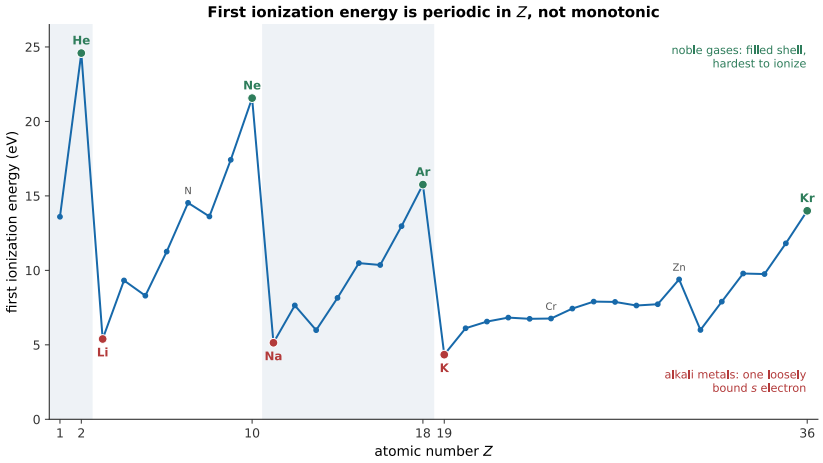

Figure 11.2: First ionization energy versus atomic number, $Z = 1$–36. Each period rises from an alkali-metal trough to a noble-gas peak as Z_{eff} grows across the period, then collapses as a new, more distant shell begins. Computed from NIST Atomic Spectra Database ionization-energy values.

Two smaller features in that figure repay a closer look. Within period 2, beryllium's ionization energy (9.32 eV) is *higher* than boron's (8.30 eV), even though boron has one more proton — a small step backward against the overall rising trend, and the same pattern recurs for nitrogen (14.53 eV) against oxygen (13.62 eV), and again in period 3 for phosphorus (10.49 eV) against sulfur (10.36 eV). These dips are not noise; they are the same subshell-stability physics responsible for the chromium and copper exceptions above, showing up in ionization energy instead of in the filling order.

Worked Example: The Ionization-Energy Dips at Boron and Oxygen

Beryllium's configuration is $1s^2 2s^2$. Removing an electron from boron ($1s^2 2s^2 2p^1$), by contrast, means removing the *single* $2p$ electron — a higher-energy, more weakly bound (less penetrating, more screened) subshell than $2s$ — which is easier despite boron's larger nuclear charge. The key is the subshell ordering produced by penetration and screening: the added boron electron

enters a $2p$ subshell that lies above $2s$, rather than a general rule that every filled subshell is unusually stable. There is also no exchange-energy advantage for boron's lone $2p$ electron comparable to the half-filled case discussed next.

Nitrogen's configuration is $1s^2 2s^2 2p^3$: a *half-filled* $2p$ subshell, with all three $2p$ electrons in separate orbitals with parallel spin, by Hund's rule. Oxygen ($1s^2 2s^2 2p^4$) must place its fourth $2p$ electron into an orbital that already holds one — pairing two electrons in the same spatial orbital, with opposite spins, for the first time in the $2p$ subshell. That pairing adds direct Coulomb repulsion between the two electrons sharing an orbital and removes some of the exchange stabilization associated with the half-filled arrangement. Both effects make oxygen's *paired* electron easier to remove than any of nitrogen's three unpaired ones, despite oxygen's larger nuclear charge — exactly the same underlying mechanism (exchange-energy stabilization of half-filled subshells) that pins chromium at $3d^5 4s^1$ rather than $3d^4 4s^2$.

11.3 Atomic Spectra, X-Rays, and Lasers

Atomic Spectra and X-Rays

Optical spectra of multi-electron atoms arise, as in hydrogen, from transitions of a single (typically outer, valence) electron between energy levels, now shifted by screening as described above and further split by the interaction between an electron's orbital and spin magnetic moments (**spin-orbit coupling**), producing the closely spaced doublets and multiplets seen in high-resolution atomic spectra.

A distinct and higher-energy class of spectral lines, **characteristic X-rays**, arises from transitions of *inner-shell* electrons. If an atom is bombarded with sufficiently energetic electrons (as in an X-ray tube) or photons, an inner-shell electron (e.g., from the $n = 1$, or K, shell) can be ejected entirely, leaving a vacancy. An electron from a higher shell then drops down to fill the vacancy, emitting a photon whose energy — since inner-shell electrons in heavier atoms feel nearly the full, largely unscreened nuclear charge Z — is far larger than typical optical-transition energies, and falls in the X-ray part of the spectrum. X-ray spectroscopists label shells by letter rather than by n: $n = 1$ is the K shell, $n = 2$ the L shell, $n = 3$ the M shell, $n = 4$ the N shell, and so on — a historical notation, predating the Bohr model, that has simply stuck. Lines from transitions ending on the K shell are called the **K series** (with K_α for the $n = 2 \to n = 1$, i.e. $L \to K$, transition, K_β for $n = 3 \to n = 1$, i.e. $M \to K$, etc.); an atom with a vacancy in the L shell instead produces the analogous, lower-energy L series (L_α for $M \to L$, and so on).

> The Greek-letter subscript counts how many shells the initial electron dropped: α for a jump from the adjacent shell ($\Delta n = 1$, e.g. $L \to K$), β for a jump from two shells up ($\Delta n = 2$, e.g. $M \to K$), and so on — the same convention used for the Lyman/Balmer-series labels of Chapter 10.

Henry Moseley (1913) measured the characteristic X-ray frequencies of many elements and found that the frequency of the K_α line follows a strikingly simple pattern, **Moseley's law**:

$$\sqrt{f} = a(Z - b), \tag{11.5}$$

where a and b are constants (with $b \approx 1$ for the K_α line, reflecting screening of the nuclear charge by the one remaining K-shell electron) essentially independent of the element. This

relation follows directly from a hydrogen-like treatment of the transition, with the true nuclear charge Z replaced by an effective charge $Z_{\text{eff}} = Z - b$ to account for screening by the other K-shell electron, applied to the Bohr/Schrödinger hydrogen energy formula. Moseley's law provided, for the first time, a direct physical (rather than merely chemical) way to determine an element's atomic number Z, and was used to correctly order elements in the periodic table, resolve several ambiguities in the ordering by atomic mass, and confirm the existence of predicted-but-then-unobserved elements by their expected X-ray frequency, cementing Z (nuclear charge) rather than atomic mass as the correct organizing quantity for the periodic table.

The Moseley Plot

Moseley's own presentation of his data was graphical rather than tabular, and remains the clearest way to see the law's content: plotting $\sqrt{f}$ (the square root of the measured K_α frequency) against Z for every element measured produces, to remarkable precision, a single **straight line**. That linearity is the entire empirical claim — a hydrogen-like $1/n^2$ energy spectrum predicts $f \propto (Z-b)^2$, i.e., $\sqrt{f} \propto (Z-b)$, so a plot linear in Z is direct evidence that inner-shell transition energies really do scale as a screened hydrogen atom's, even in atoms far too complicated to solve exactly. The same linear relationship holds separately for the L series and other series, each with its own slope a and intercept b (the L-series screening constant b is larger than the K-series value of about 1, since an L-shell vacancy is screened by more surrounding electrons than a K-shell vacancy is) — a family of parallel diagnostic lines, one per series, any one of which pins down Z from a measured frequency alone.

Worked Example: Moseley's Law — Predicting Nickel's K_α Line from Copper's

A single measured K_α frequency, together with the known screening constant $b \approx 1$, is enough to predict the K_α frequency of a neighboring element. Copper ($Z = 29$) has a measured K_α frequency $f_{\text{Cu}} = 1.94 \times 10^{18}$ Hz (corresponding to a photon energy $hf_{\text{Cu}} \approx 8.02$ keV, close to the accepted value of 8.048 keV for the Cu $K_{\alpha 1}$ line). Writing Moseley's law as $\sqrt{f} = A(Z-1)$ and solving for A using copper's data,

$$A = \frac{\sqrt{f_{\text{Cu}}}}{29-1} = \frac{\sqrt{1.94 \times 10^{18}\ \text{Hz}}}{28} = \frac{1.393 \times 10^{9}\ \sqrt{\text{Hz}}}{28} = 4.974 \times 10^{7}\ \sqrt{\text{Hz}}. \tag{11.6}$$

Applying the same A to nickel ($Z = 28$, one proton lighter),

$$\sqrt{f_{\text{Ni}}} = A(28-1) = (4.974\times10^{7}\ \sqrt{\text{Hz}})(27) = 1.343\times10^{9}\ \sqrt{\text{Hz}}, \qquad f_{\text{Ni}} \approx 1.80\times10^{18}\ \text{Hz}, \tag{11.7}$$

corresponding to a predicted photon energy $hf_{\text{Ni}} \approx 7.46$ keV. The accepted experimental value for the Ni K_α doublet is 7.461–7.478 keV — agreement to better than half a percent, from a calculation using only a single measured data point and one universal constant, exactly the kind of predictive power that let Moseley assign atomic numbers to elements from X-ray measurements alone.

The surviving photograph in Figure 11.3 shows the physicist himself, not long before the work described above.

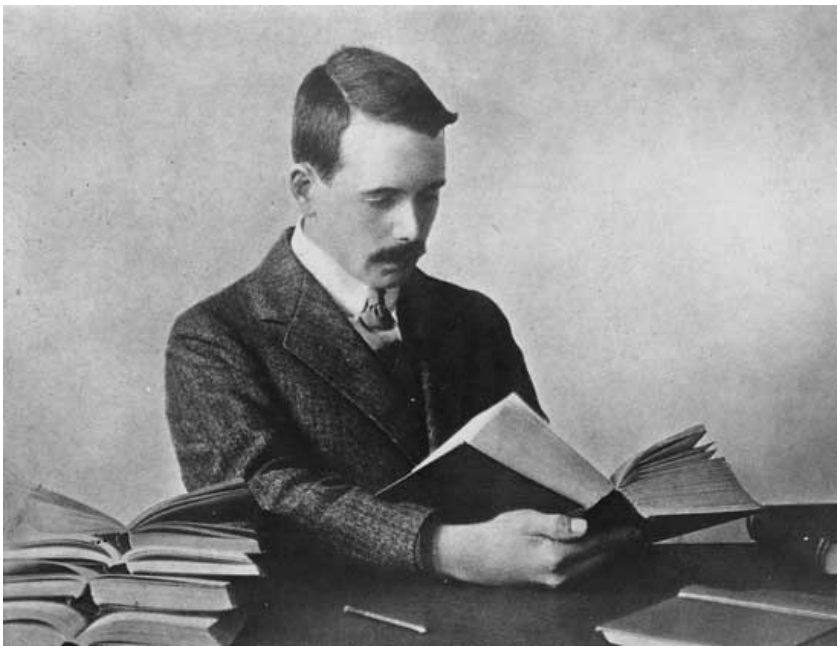

Figure 11.3: Henry Gwyn Jeffreys Moseley, circa 1914. Photograph via the AIP Emilio Segrè Visual Archives, W. F. Meggers Gallery of Nobel Laureates; public domain via Wikimedia Commons.

Historical Context: Henry Moseley and the 1913–1914 X-Ray Survey

Henry Moseley was 26 years old, working at Ernest Rutherford's laboratory in Manchester and then at Oxford, when he built an X-ray spectrometer and, over roughly a year, measured the characteristic K- and L-series frequencies of nearly 40 elements — an extraordinarily fast and thorough experimental survey by the standards of the day. The result was immediately and directly useful to chemistry: several elements were known, from their chemical properties, to sit in a different order than their measured atomic *masses* would suggest (tellurium and iodine are the classic example — tellurium is heavier but chemically belongs *before* iodine in the periodic table), and Moseley's X-ray frequencies confirmed that atomic *number*, not mass, was the correct ordering quantity in every such case. His data also revealed gaps: no element then known produced the K_α frequency predicted for $Z = 43$, 61, or 75, and Moseley correctly concluded that these elements simply had not yet been discovered (they are now known as technetium, promethium, and rhenium; technetium and promethium are highly radioactive and occur only in trace amounts naturally. Technetium was identified in 1937 after being produced from a cyclotron target, although trace natural technetium does exist). A vacancy at $Z = 72$ was, for a decade, the subject of a bitter chemical priority dispute — a French chemist had already claimed the element among the rare earths and named it "celtium" — until 1922, when X-ray measurements at Bohr's institute in Copenhagen showed conclusively that the true element 72 (hafnium) is chemically a heavier relative of zirconium, not a rare earth at all, and had simply been misidentified.

Moseley enlisted in the Royal Engineers at the outbreak of the First World War and was killed at Gallipoli on 10 August 1915, at 27, shot through the head while relaying a message during the landing. His death is often cited (including by later Nobel laureates who worked to change the practice) as a direct argument for the British and other governments' subsequent policy of not sending scientifically valuable personnel to serve at the front line — a policy adopted too late to save Moseley himself, whose X-ray survey had already, in barely two years,

put the periodic table's underlying organizing principle beyond doubt.

At the optical end of the same physics, Figure 11.4 runs the experiment a neon sign runs: electrons accelerated through a low-pressure gas, colliding with atoms, and the light that comes out. The collisions are the point. Below the threshold for inelastic excitation, collisions are mostly elastic and the tube stays dark; raise the accelerating voltage past the first excitation energy and electrons can transfer that discrete excitation energy to the atoms, retaining the rest of their kinetic energy. The excited atoms subsequently emit photons at wavelengths fixed by their allowed transitions, sometimes through cascades of intermediate states. That threshold behavior is the Franck–Hertz experiment, and it is direct evidence of discrete atomic levels that owes nothing to spectroscopy. Multi-level atoms can be selected to see cascades through intermediate states, which is the same level structure the next section exploits.

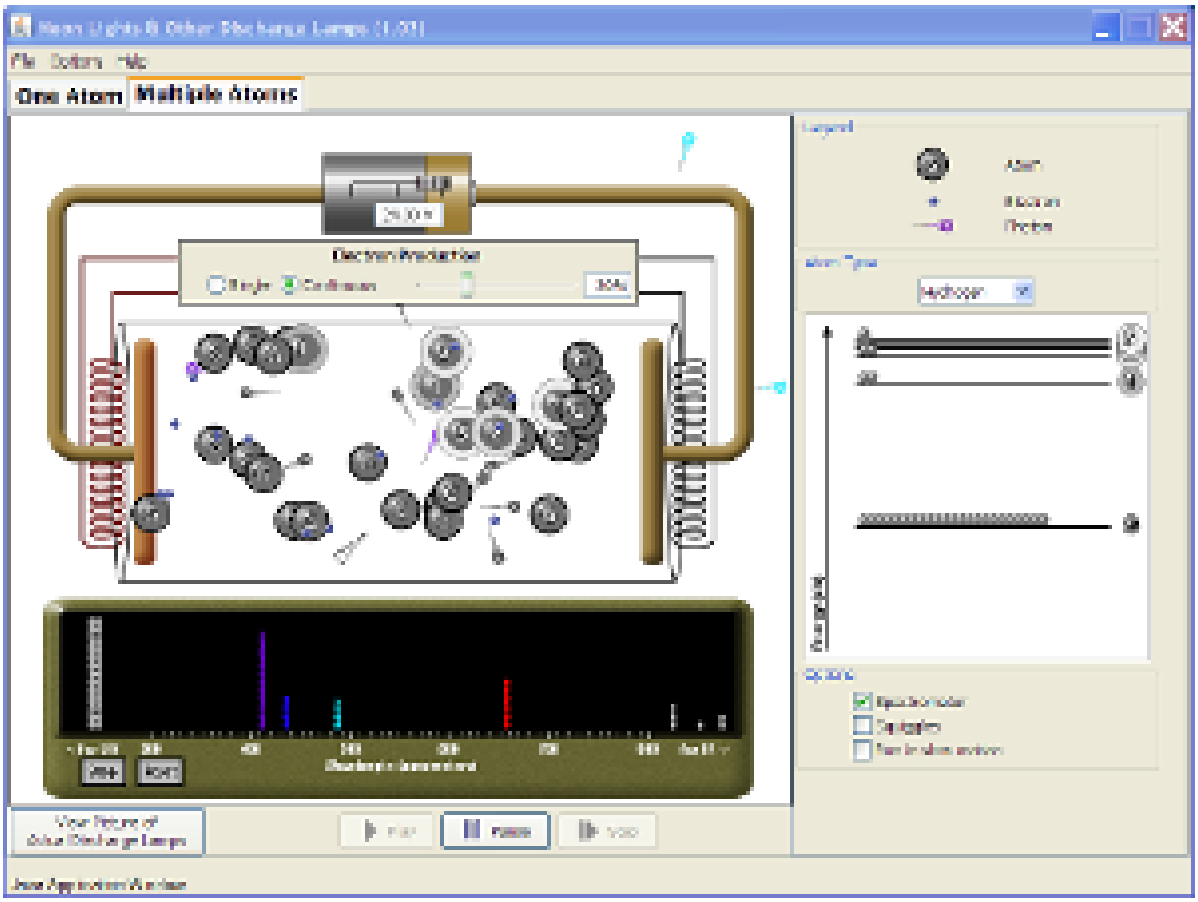

Figure 11.4: A gas discharge tube with the accelerating voltage, the gas, and the atomic energy-level scheme under control. Excitation happens only at discrete electron energies, and the emitted spectrum is the level diagram read out in light. Interactive simulation: Neon Lights & Other Discharge Lamps

Lasers

A further application of atomic energy levels involves the interaction between atoms and light more actively than simple absorption/emission spectroscopy. An atom in an excited state can lose energy in two distinct ways: **spontaneous emission**, in which it decays at a random time with a photon emitted in a random direction (governing ordinary fluorescence and, statistically, the exponential decay laws seen throughout atomic and nuclear physics), and **stimulated**

emission, in which a passing photon of exactly the transition energy triggers the atom to emit a second, additional photon that is an exact copy of the first — same energy, direction, phase, and polarization. Einstein first predicted stimulated emission in 1917, well before it could be technologically exploited.

Ordinarily, stimulated emission is masked by the competing process of absorption, since a typical collection of atoms in thermal equilibrium has more atoms in lower-energy states than higher ones. A **laser** (light amplification by stimulated emission of radiation) requires engineering a **population inversion** — an atomic sample with more atoms in a higher-energy state than a lower one, typically achieved by "pumping" atoms into a higher state and relying on an intermediate **metastable state** (one with an anomalously long lifetime against spontaneous decay, because its decay to lower states is forbidden or strongly suppressed by selection rules such as the $\Delta\ell = \pm 1$ rule of Chapter 10) to accumulate a population large enough to exceed that of the lower lasing level. Once inverted, a single spontaneously emitted photon can trigger a cascade of stimulated emission as it passes back and forth through the gain medium (typically between two mirrors forming an optical cavity), producing an intense, coherent, highly directional, single-wavelength beam — a direct, large-scale, technological manifestation of discrete atomic energy levels and the quantum-mechanical description of light-matter interaction developed across this and the preceding chapters.

Two Concrete Laser Systems: Helium–Neon and Ruby

The **helium–neon (He-Ne) laser**, one of the first lasers built (1961) and still common in classrooms and barcode scanners, illustrates a **four-level scheme**. An electrical discharge through a low-pressure He-Ne gas mixture excites helium atoms into a long-lived metastable state at 20.61 eV above the helium ground state — an s-state whose direct decay to the (also s-state) helium ground level is forbidden by the $\Delta\ell = \pm 1$ selection rule, not merely unlikely because of its energy, so helium atoms accumulate there in large numbers, exactly the metastable-state bottleneck described above. That energy happens to very nearly match a neon excited state near 20.66 eV, so an excited helium atom colliding with a ground-state neon atom can transfer its energy directly, exciting the neon (and leaving the helium atom free to be re-excited by the discharge) — helium serves purely as an energy-transfer intermediary, never itself the lasing species. The excited neon atom then undergoes stimulated emission down to a lower excited state, emitting the familiar 632.8 nm red beam; because that lower level is itself well above the neon ground state and decays away rapidly through further (non-lasing) transitions, it never accumulates a large population, so the population inversion between the upper and lower laser levels is easy to sustain — the defining advantage of a four-level design over a three-level one.

The **ruby laser**, built by Theodore Maiman in 1960 (the first laser of any kind to operate), is a **three-level system** and correspondingly harder to pump. Ruby is aluminum oxide doped with a small fraction of chromium ions (Cr^{3+}), whose electronic transitions (not those of the aluminum oxide host) do the lasing. An intense flash lamp pumps chromium ions from the ground state into two broad, short-lived absorption bands (centered in the blue and green, which is why ruby appears red — it absorbs the complementary colors); those ions decay almost immediately, without emitting a photon, to a metastable level about 1.79 eV above the ground state, with a comparatively long lifetime of several milliseconds. Because the *lower* laser level in this scheme is the atomic ground state itself, which starts out fully populated, achieving a population inversion requires pumping *more than half* of all the chromium ions into the metastable state simultaneously — a substantially harder threshold to cross than in a four-level

scheme, where the lower laser level is nearly empty to begin with. Once inverted, stimulated emission from the metastable level back to the ground state produces ruby's characteristic 694.3 nm deep-red beam, historically in short, intense pulses rather than the continuous output a He-Ne laser can sustain.

Every ingredient of that description is separately adjustable in Figure 11.5, which is the best way to see that a laser is a threshold device rather than a bright lamp. Pump a two-level medium as hard as you like and it will not lase: absorption and stimulated emission from a two-level system saturate at equal populations, so no inversion is possible. Switch to three levels, so that atoms accumulate in a metastable state, and the inversion appears; add the cavity mirrors and, above a pump rate that the simulation makes it easy to bracket from below, the output collapses from a random scatter of spontaneous photons into a single coherent beam.

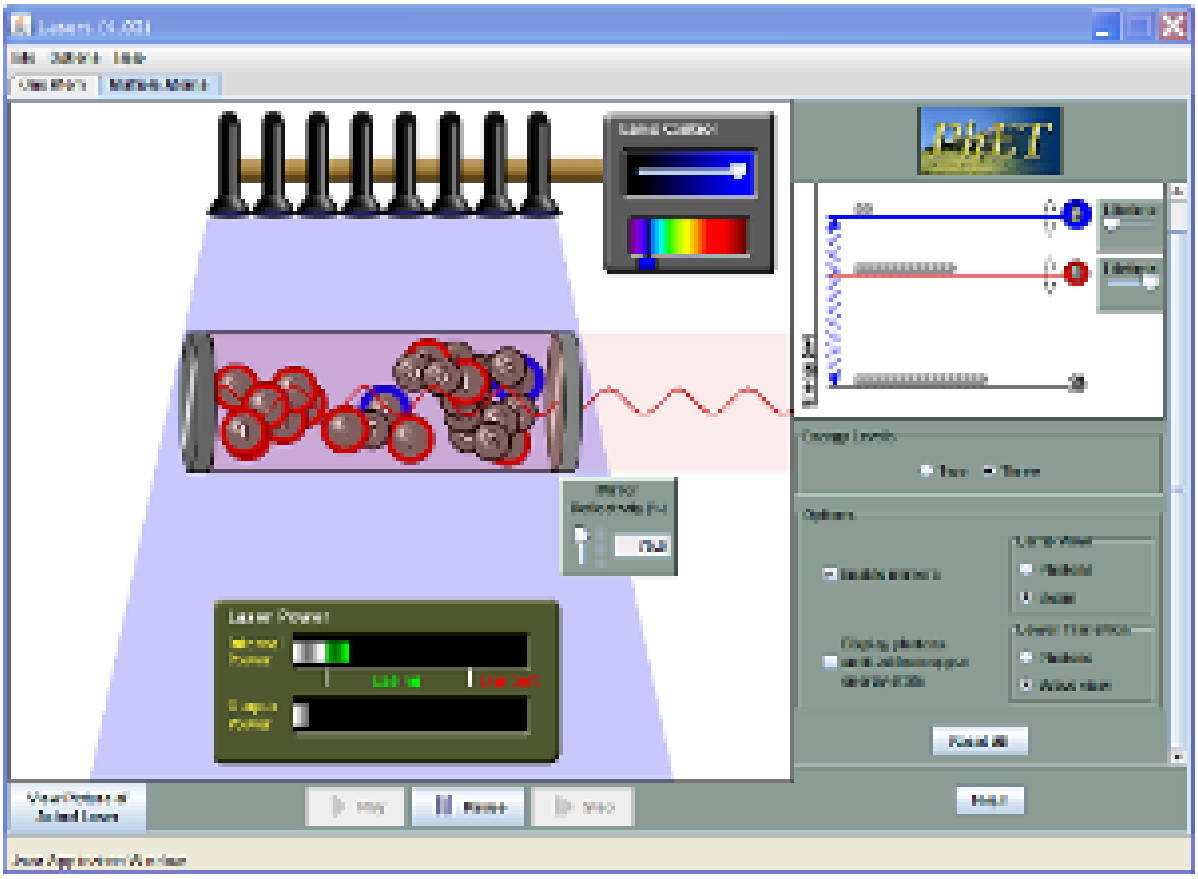

Figure 11.5: A laser assembled from its parts: a pump, a two- or three-level gain medium, and an optical cavity. Population inversion is displayed directly, so the threshold for lasing can be found by experiment rather than asserted. Interactive simulation: Lasers

Worked Example: Population Inversion Requires Active Pumping

It is worth confirming, numerically, why a population inversion never arises on its own from thermal equilibrium and must be engineered by pumping. In thermal equilibrium at temperature T, the ratio of the number of atoms in an upper state (energy E_2) to a lower state (energy E_1) is set by the Boltzmann factor,

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{k_B T}\right). \tag{11.8}$$

For the ruby laser transition, $E_2 - E_1 \approx 1.79$ eV (found from hc/λ with $\lambda = 694.3$ nm, using $hc = 1240\ \text{eV} \cdot \text{nm}$). At room temperature, $T \approx 300$ K, $k_B T \approx 0.02585$ eV, so

$$\frac{N_2}{N_1} = \exp\left(-\frac{1.79\ \text{eV}}{0.02585\ \text{eV}}\right) = \exp(-69.2) \approx 8 \times 10^{-31}. \tag{11.9}$$

Out of every 10^{30} or so chromium ions in thermal equilibrium, essentially none sit in the upper laser level — an inversion ($N_2 > N_1$) is not just unlikely but is astronomically, hopelessly far from anything thermal equilibrium could ever produce at ordinary temperature. This is precisely why a laser requires an external pump (an intense flash lamp, an electrical discharge, or another laser) to force the population far out of thermal equilibrium, and why, once the pump is switched off, the inverted population decays and lasing stops — a laser is fundamentally a *driven*, non-equilibrium device, not a system that has simply been heated.

Applications: Lasers in Technology and Medicine

The properties that make a laser beam distinctive — a single, precisely defined wavelength; a high degree of spatial coherence, allowing the beam to be focused to a tiny spot or to travel long distances with little spreading; and a high degree of temporal coherence, permitting long coherence lengths (Chapter 4) — underlie applications far removed from the gas-discharge tubes and ruby rods described above. **Fiber-optic communication**, the physical backbone of the modern internet, encodes digital data as pulses from semiconductor **diode lasers** (a solid-state laser design, distinct from both examples above, in which the population inversion is created electrically across a semiconductor junction rather than by an optical pump) sent down thin glass fibers by total internal reflection; the laser's narrow wavelength spread keeps different colors of light traveling through the same fiber from smearing into one another over long distances, allowing enormous data rates over transoceanic distances. **Laser cooling**, by contrast, exploits the momentum $p = h/\lambda$ carried by individual photons (Chapter 6): a beam tuned just below an atom's resonant absorption frequency preferentially scatters off atoms moving *toward* the beam (Doppler-shifted into resonance), each absorption event imparting a momentum kick that opposes the atom's motion, and repeated scattering from beams in multiple directions can cool a dilute atomic gas to temperatures of a microkelvin or below — a technique essential to modern atomic clocks and to the experimental study of quantum gases. In medicine, lasers are used both destructively, where a tightly focused, high-intensity beam cuts or cauterizes tissue with a precision unavailable to a mechanical scalpel (as in some ophthalmic and dermatological surgery), and diagnostically, where the coherence and narrow linewidth of laser light are exploited in optical coherence tomography to image tissue structure with micrometer resolution — different laser properties from the list above put to work in each case.

11.4 Summary

- The **Pauli exclusion principle** forbids two electrons from sharing the same full set of quantum numbers (n, ℓ, m_ℓ, m_s), limiting each orbital to two electrons (opposite spin) and is essential to explaining atomic structure.

- **Screening** of the nuclear charge by inner electrons makes subshell energy depend on both n and ℓ in multi-electron atoms, unlike hydrogen; lower-ℓ orbitals penetrate closer to the nucleus and lie lower in energy at fixed n. The **self-consistent field (Hartree) method** computes screening explicitly, by iterating each electron's wave function in the averaged potential of all the others until the potential stops changing; **Slater's rules** give a fast hand-calculable estimate of the resulting effective nuclear charge Z_{eff}.
- The **Aufbau principle** (fill lowest-energy subshells first, respecting exclusion) and **Hund's rule** (maximize total spin within a partially filled subshell) determine ground-state electron configurations, which in turn generate the row/column/block structure of the periodic table and the periodicity of chemical properties. **Exchange and subshell-closure effects** favor especially stable half-filled and completely filled subshells, producing Aufbau exceptions (chromium, copper) and small dips in ionization energy (boron, oxygen) that the naive filling order alone does not predict.
- **First ionization energy** generally rises across a period while **atomic radius** generally falls, both tracking Z_{eff}; down a group, the trends reverse as increasing n dominates a nearly constant Z_{eff}.
- **Characteristic X-rays** arise from inner-shell vacancies, labeled by shell letter (K: $n = 1$, L: $n = 2$, M: $n = 3$, …); **Moseley's law**, $\sqrt{f} = a(Z - b)$ — a straight line when $\sqrt{f}$ is plotted against Z — let X-ray spectra be used to determine atomic number directly, correctly ordering the periodic table by Z and revealing the undiscovered elements $Z = 43$, 61, and 75.
- **Lasers** exploit stimulated emission and a **population inversion**, sustained via a **metastable state**, to produce coherent light; three-level (ruby) and four-level (helium–neon) schemes differ in how hard the lower laser level is to keep empty. Lasers underlie fiber-optic communication, laser cooling, and a range of medical applications.

11.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens on the periodic ionization-energy trend.

1. **Multiple choice, from a figure.** What does screening do to the nuclear attraction felt by an outer electron? Figure description: First ionization energy plotted across the periodic table, showing periodic trends. Choices: (A) It reduces the effective nuclear charge felt by the electron. (B) It reverses the nuclear force and makes it repulsive. (C) It removes all dependence on orbital penetration. (D) It forces every subshell with the same n to remain degenerate.

2. **True or false.** Two electrons in one spatial orbital may have the same spin projection.

3. **Drag the words.** Complete the rules for many-electron atoms. The _____ limits an orbital to two opposite-spin electrons. The _____ fills lower-energy subshells first. _____ maximizes spin in degenerate orbitals. Screening produces an _____. (Terms: Pauli exclusion principle; Aufbau principle; Hund's rule; effective nuclear charge; uncertainty principle; bare nuclear charge for every electron.)

4. **Fill in the blanks.** Complete the shell and laser statements. A K-shell vacancy lies in the shell with n = _____, and laser amplification requires a population _____.

5. **Mark the words.** Mark the two general trends across a period. "Across a period, first ionization energy generally increases while atomic radius generally decreases, as effective nuclear charge grows."

11.6 Conceptual Questions

1. How does the exclusion principle generate shell structure without introducing a new force between electrons?
2. Why are $4s$ orbitals filled before $3d$ in neutral atoms but often ionized first in transition-metal ions?
3. Explain Hund's rule in terms of how electrons occupy degenerate orbitals.
4. Why do atomic radii generally decrease across a period even as more electrons are added?
5. How do characteristic X-rays reveal the energy differences between inner electron shells?
6. Distinguish spontaneous emission, stimulated emission, and absorption.

11.7 Problems

Exercise 11.1
(Straightforward)

Write the ground-state electron configuration (using $n\ell^x$ notation, e.g. $1s^2\,2s^2\ldots$) for (a) carbon ($Z = 6$), (b) sodium ($Z = 11$), (c) iron ($Z = 26$), using the filling order given in the text.

Exercise 11.2
(Straightforward)

Using Hund's rule, sketch the orbital-filling diagram (boxes for each m_ℓ orbital, arrows for spin) for the $2p$ subshell of nitrogen ($Z = 7$, configuration $1s^2 2s^2 2p^3$), and state the resulting total spin.

Exercise 11.3
(Straightforward)

Explain why, in a multi-electron atom, a $4s$ electron can have lower energy than a $3d$ electron despite having a larger principal quantum number, using the concept of orbital penetration and screening.

Exercise 11.4
(Straightforward)

The measured K_α X-ray frequency of copper ($Z = 29$) is $f = 1.94 \times 10^{18}$ Hz. Using Moseley's law in the form $\sqrt{f} = A(Z - 1)$ (i.e., $b = 1$) with a single data point to determine A, predict the K_α frequency of nickel ($Z = 28$), and compare qualitatively to what you would expect (higher or lower than copper's).

Exercise 11.5
(Moderate)

Explain, in terms of the exclusion principle, why the ground-state electron configuration of helium ($1s^2$) is chemically inert, while lithium ($1s^2 2s^1$) is highly reactive, referring to the energy required to remove the outermost electron in each case.

Exercise 11.6
(Moderate)

Explain why a three-level or four-level laser scheme requires a metastable intermediate state to sustain a population inversion, rather than pumping directly into the lower lasing level's excited partner state, using the relative decay rates implied by allowed versus forbidden transitions (Chapter 10's selection rule).

Exercise 11.7
(Moderate)

Using Slater's rules (the grouping $(1s)(2s,2p)(3s,3p)\ldots$ and the shielding constants 0.35/same group, 0.85/one shell lower, 1.00/two or more shells lower), compute the effective nuclear charge Z_{eff} felt by a $3s$ electron in magnesium ($Z = 12$, configuration $1s^2 2s^2 2p^6 3s^2$). Compare your result to the text's values of $Z_{\text{eff}} = 2.20$ for sodium's $3s$ electron and $Z_{\text{eff}} = 6.10$ for chlorine's $3p$ electron, and state whether magnesium's value is consistent with its position between them in the periodic table.

Exercise 11.8
(Moderate)

Sulfur's first ionization energy (10.36 eV) is slightly *lower* than phosphorus's (10.49 eV), even though sulfur has one more proton. Using phosphorus's and sulfur's electron configurations ($[\text{Ne}]\,3s^2 3p^3$ and $[\text{Ne}]\,3s^2 3p^4$, respectively) and the exchange-energy argument given in the text for the analogous nitrogen/oxygen dip, explain this reversal.

Exercise 11.9
(Moderate)

Copper's K_α frequency is $f_{\text{Cu}} = 1.94 \times 10^{18}$ Hz ($Z = 29$); molybdenum's measured K_α photon energy is 17.478 keV ($Z = 42$). (a) Using both data points in Moseley's law $\sqrt{f} = a(Z - b)$ (two equations, two unknowns a and b), solve for a and b, and compare b to the approximate value $b \approx 1$ quoted in the text for the K_α line. (b) Use your fitted a and b to predict the K_α photon energy of silver ($Z = 47$).

Exercise 11.10
(Challenging)

A helium–neon laser emits a continuous beam at $\lambda = 632.8$ nm with an output power of 5.00 mW. (a) Find the energy of a single photon at this wavelength, in eV and in joules. (b) Find the number of photons emitted per second. (c) Explain briefly why this beam, despite its low power compared to an ordinary light bulb, can still be hazardous to the retina, referring to the spatial coherence discussed in the text.

Exercise 11.11

(Challenging)

In your own words, explain why the Hartree self-consistent field method is "circular" — that is, why the effective potential $V_{\text{eff}}(r)$ needed to solve for an electron's wave function cannot simply be written down in advance — and describe the iterative procedure used to resolve that circularity, including what "self-consistent" means as a stopping criterion.

Exercise 11.12

(Challenging)

Argon ($Z = 18$, configuration $[\text{Ne}]\,3s^2 3p^6$) has a *smaller* atomic radius than the very next element, potassium ($Z = 19$, configuration $[\text{Ar}]\,4s^1$), even though potassium has one more proton and one more electron. Using the concepts of principal quantum number n and effective nuclear charge Z_{eff} developed in this chapter, explain why adding a proton and an electron can *increase* atomic radius in this specific case, when it decreases atomic radius everywhere else within period 3.

Chapter 12

Molecular Structure

Learning Objectives

By the end of this chapter, you should be able to:

- Distinguish ionic and covalent bonding in terms of the underlying electron distribution.
- Explain covalent bond formation in valence bond theory as the overlap of atomic orbitals.
- Explain the need for hybrid orbitals and identify the hybridization and molecular geometry implied by a given number of electron domains.
- Construct molecular orbitals as linear combinations of atomic orbitals and distinguish bonding from antibonding orbitals.
- Use a molecular orbital diagram to compute bond order and predict the stability and magnetic behavior of a simple diatomic molecule.
- Explain the vibrational and rotational energy levels of a diatomic molecule using the harmonic-oscillator and rigid-rotor approximations.

Introduction

Chapters 10 and 11 explained the structure of individual atoms — how electrons occupy discrete energy levels arranged into subshells and shells, and how that arrangement produces the periodic table. This chapter asks how atoms combine to form molecules, using the same quantum-mechanical toolkit: atomic orbitals, the Pauli exclusion principle, and the variational tendency of a bound system to seek its lowest-energy configuration. Two complementary pictures are developed. **Valence bond theory** treats a bond as the overlap of atomic orbitals from two atoms, localized between them, and is the natural language for molecular geometry. **Molecular orbital theory** instead builds orbitals belonging to the molecule as a whole, and is the more powerful tool for predicting a molecule's stability, bond strength, and magnetic properties. The chapter closes by treating a bonded diatomic molecule as a single quantum system in its own right, subject to quantized vibrational and rotational energy levels — a direct application of the harmonic oscillator (Chapter 8) and angular momentum quantization (Chapter 9) to a new physical system.

12.1 Bonding and Hybrid Orbitals

Ionic and Covalent Bonding

Chemical bonds form because a molecule can have lower total energy than its constituent separated atoms. Two limiting mechanisms produce this energy lowering. In **ionic bonding**, one atom (typically one with a low ionization energy, such as an alkali metal) transfers one or more electrons entirely to another atom (typically one with a high electron affinity, such as a halogen); the resulting oppositely charged ions are then held together by simple electrostatic (Coulomb) attraction. In **covalent bonding**, by contrast, one or more electron pairs are *shared* between two atoms, occupying a region of enhanced electron density between the two nuclei; both nuclei are then simultaneously attracted to this shared, concentrated negative charge, producing a net attractive bond. Most real bonds fall on a continuum between these two limits, described by varying degrees of bond **polarity**, depending on the difference in electronegativity between the bonded atoms; this chapter focuses on the covalent limit, whose treatment requires genuinely quantum-mechanical ideas beyond simple electrostatics.

> **Electronegativity**, on the dimensionless Pauling scale, runs from about 0.7 (cesium) to 4.0 (fluorine). A difference of roughly 1.7 or more between two bonded atoms is the conventional (rough, not sharp) threshold above which a bond is usually described as "ionic" rather than "polar covalent."

That continuum is a slider rather than a classification, and Figure 12.1 makes it one: set the electronegativity of each atom in a diatomic and watch the shared electron density slide from the midpoint toward the more electronegative partner, dragging a dipole moment with it. At equal electronegativity the bond is purely covalent and the dipole vanishes; at a large difference the electron is effectively transferred and what remains is a pair of ions attracting each other. Neither limit is a separate mechanism — they are the two ends of one.

Valence Bond Theory and Orbital Overlap

Valence bond theory treats covalent bond formation as arising from the overlap of a singly occupied atomic orbital on one atom with a singly occupied atomic orbital on another, the two electrons (one from each atom, necessarily of opposite spin, per the exclusion principle applied to the resulting shared, doubly occupied region) pairing up to form the bond. The simplest example is the hydrogen molecule H_2: as two hydrogen atoms approach, their $1s$ orbitals begin to overlap, and if the two electrons involved have opposite spin, the resulting overlap region between the nuclei has a high joint probability of finding both electrons — an enhanced electron density that lowers the system's total energy relative to two separate atoms, up to a certain optimal internuclear separation (the **bond length**, at which attractive and repulsive contributions to the energy balance). Bonds formed by orbitals overlapping directly along the internuclear axis, giving a cylindrically symmetric electron distribution about that axis, are called **sigma (σ) bonds**; bonds formed by the sideways overlap of parallel p orbitals, with electron density concentrated above and below (rather than directly along) the internuclear axis, are called **pi (π) bonds**. A single bond is one σ bond; a double bond is one σ plus one π bond; a triple bond is one σ plus two (mutually perpendicular) π bonds.

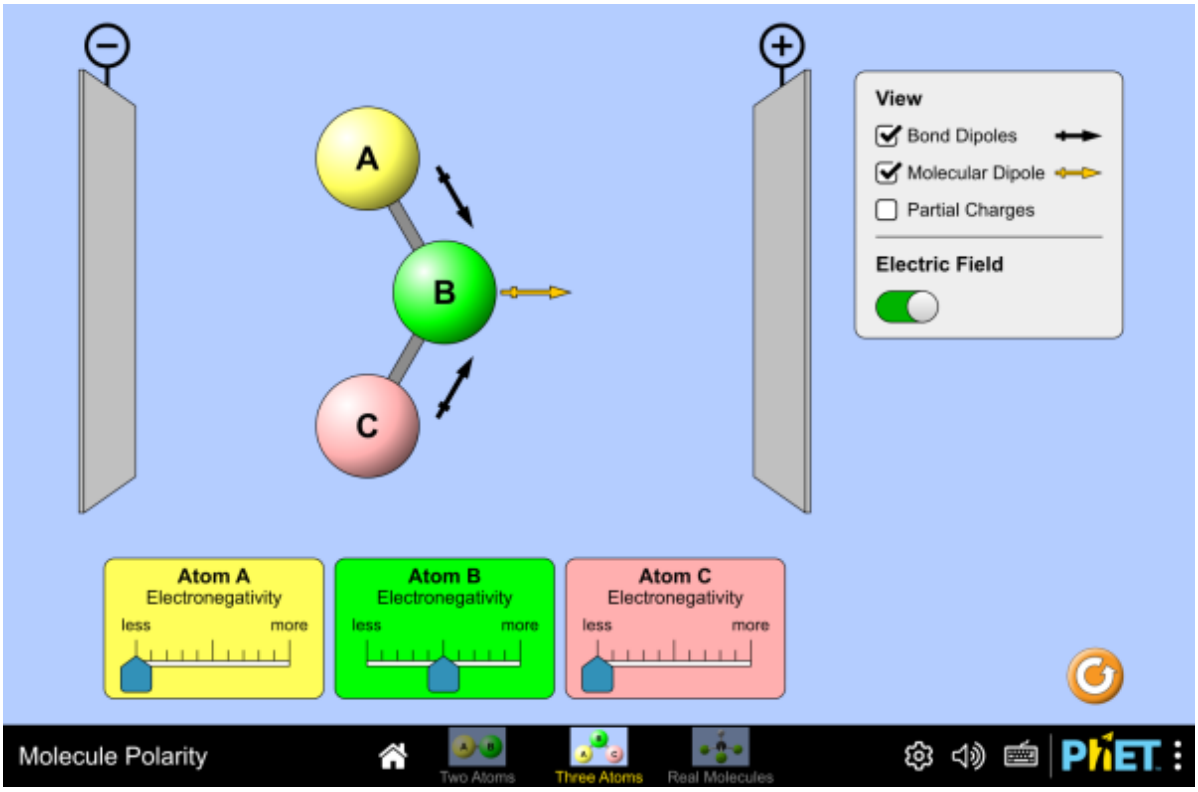

Figure 12.1: Bond polarity as a continuous function of the electronegativity difference between two atoms, with the electron density and the resulting dipole moment displayed. Later screens add a third atom, where bond dipoles combine vectorially and a molecule of polar bonds can still be nonpolar overall. Interactive simulation: Molecule Polarity

Table 12.1: Common orbital hybridizations and their molecular geometries

Hybridization	Orbitals mixed	Number of hybrids	Geometry	Example
sp	one s, one p	2	linear (180°)	$BeCl_2$
sp^2	one s, two p	3	trigonal planar (120°)	BF_3
sp^3	one s, three p	4	tetrahedral (109.5°)	CH_4
sp^3d	one s, three p, one d	5	trigonal bipyramidal	PCl_5
sp^3d^2	one s, three p, two d	6	octahedral	SF_6

The labels σ and π describe the *symmetry* of the orbital overlap about the bond axis, not the bond's strength or order directly: a single bond is always exactly one σ bond, however the orbitals forming it are hybridized, and every bond beyond the first in a double or triple bond is necessarily a π bond.

Hybrid Orbitals

Simple atomic s and p orbitals, taken directly from the hydrogen-like solutions of Chapter 10, generally point in the wrong directions (or have the wrong shapes) to account for the observed bond angles and geometries of real molecules — the observed near-109.5° bond angles of methane, CH_4, for instance, are not reproduced by combinations of the atom's unmodified $2s$ and three $2p$ orbitals directly. The resolution is that the atomic orbitals actually used in bonding are not the pure s, p, d orbitals of an isolated atom, but specific linear combinations of them, called **hybrid orbitals**, that better match the geometry demanded by minimizing electron-pair repulsion among the atom's bonding and lone electron pairs (the same qualitative principle underlying the electron-domain, or VSEPR, geometries encountered in general chemistry). Mixing one s orbital with a varying number of p (and, for some geometries, d) orbitals produces hybrid sets with characteristic, experimentally matched geometries:

The standard geometries are collected in Table 12.1.

For the five- and six-domain cases, sp^3d and sp^3d^2 are traditional hybridization labels that summarize the electron-domain geometry. They should not be read as evidence that occupied valence d orbitals are literally needed to form the bonds in hypervalent molecules such as PCl_5 and SF_6; modern bonding descriptions account for these molecules without substantial d-orbital participation.

The general rule connecting geometry to hybridization is that the number of hybrid orbitals equals the number of **electron domains** around the central atom — bonding pairs plus lone pairs — and that hybrid orbitals arrange themselves to maximize their mutual angular separation, minimizing electron-pair repulsion, exactly as in the VSEPR (valence-shell electron-pair repulsion) model of molecular geometry. Unshared (lone) electron pairs occupy hybrid orbitals just as bonding pairs do, but exert somewhat greater repulsion (being attracted to only one nucleus rather than shared between two), which is why, for example, the bond angle in water (H_2O, two bonding pairs and two lone pairs on an approximately sp^3 oxygen) is

compressed to about 104.5° from the ideal tetrahedral 109.5°.

The counting rule behind Table 12.1 is worth exercising on molecules other than the five listed, and Figure 12.2 is set up for exactly that: attach single, double, or triple bonds and lone pairs to a central atom in any combination and the geometry rearranges itself, with the bond angles reported as they change. Two predictions from the paragraph above can be checked in a few seconds there. A double bond and a single bond count the same for geometry — both are one electron domain — and replacing a bonding pair by a lone pair squeezes the remaining angles, reproducing water's 104.5° from tetrahedral.

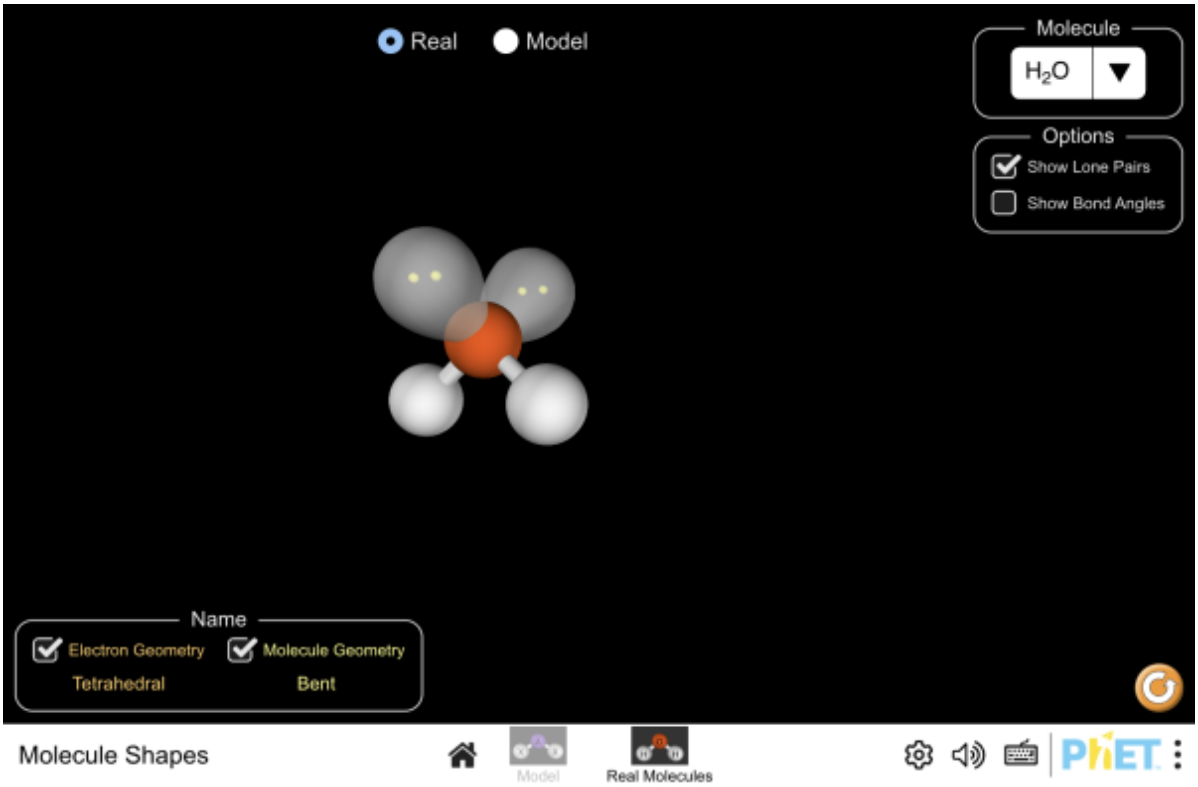

Figure 12.2: Molecular geometry built from electron domains. Bond angles are displayed live, and the "real molecules" screen compares the idealized VSEPR prediction with measured geometries. Interactive simulation: Molecule Shapes

Worked Example: The Geometry of Xenon Tetrafluoride

Predict the hybridization and molecular geometry of XeF_4.

Xenon forms four Xe–F σ bonds to the four fluorine atoms. After those four bonding pairs and the three lone pairs required on each fluorine to complete its own octet are accounted for, the count of xenon's eight valence electrons leaves **two lone pairs on xenon itself** — xenon, unlike carbon or nitrogen, has enough valence electrons to bond fully *and* retain lone pairs of its own. The central atom therefore has 4 bonding domains + 2 lone pairs = 6 electron domains, calling for sp^3d^2 **hybridization** and an octahedral arrangement of those six domains.

The two lone pairs could sit at 90° (adjacent) or 180° (opposite) to each other on the octahedron. Because lone-pair–lone-pair repulsion is the strongest of the three repulsion types recognized by the electron-domain model (lone pair–lone pair > lone pair–bonding pair > bonding pair–bonding pair), the two lone pairs move as far apart as the octahedral geometry

allows: directly opposite one another. That leaves the four Xe–F bonds occupying the remaining four positions, all in a single plane — a **square planar** molecular geometry, with F–Xe–F bond angles of exactly 90° between adjacent fluorines and 180° across. This result is worth remembering precisely because the naive guess for "four bonded groups" — tetrahedral, as in CH_4 — is wrong here: the two lone pairs, invisible in a skeletal formula, are what flatten the molecule into a plane.

12.2 Molecular Orbital Theory

Valence bond theory, with hybridization added, accounts well for molecular geometry, but it treats bonding electrons as localized between two specific atoms and struggles to describe phenomena in which electrons are shared more broadly, or where a simple bonding picture predicts the wrong number of unpaired electrons. **Molecular orbital (MO) theory** instead constructs orbitals belonging to the molecule as a whole, built as linear combinations of the atomic orbitals (LCAO) of the constituent atoms, exactly as a molecular wave function must ultimately be some solution of the (approximate, many-electron) Schrödinger equation for the whole molecule.

For two hydrogen $1s$ orbitals, ψ_A and ψ_B, on atoms A and B, the two independent linear combinations are

$$\psi_{\mathrm{MO}}^{\pm} = \psi_A \pm \psi_B. \tag{12.1}$$

The symmetric combination, $\psi_{\mathrm{MO}}^{+} = \psi_A + \psi_B$, adds constructively in the region between the two nuclei, producing enhanced electron density there and a lower energy than the separate atomic orbitals — a **bonding orbital**, denoted σ_{1s}. The antisymmetric combination, $\psi_{\mathrm{MO}}^{-} = \psi_A - \psi_B$, has a node exactly at the midpoint between the nuclei, *depleting* electron density in the internuclear region and yielding a *higher* energy than the separate atomic orbitals — an **antibonding orbital**, denoted σ_{1s}^{*}. In general, combining N atomic orbitals produces exactly N molecular orbitals (never more, never fewer) — a direct consequence of treating the LCAO expansion as a change of basis for the same underlying space of trial wave functions. The resulting orbitals can be bonding, antibonding, or, when an orbital has little net interaction with the others, nonbonding; the bonding and antibonding energies need not be symmetrically displaced from the original atomic-orbital energy.

Why the Antibonding Orbital Rises More Than the Bonding Orbital Falls

The qualitative claim that ψ_{MO}^{+} is lower in energy and ψ_{MO}^{-} is higher can be made quantitative. Define three integrals: the **overlap integral** $S = \int \psi_A \psi_B \, d\tau$ (how much the two atomic orbitals overlap in space, $0 \le S \le 1$), the **Coulomb integral** $\alpha = \int \psi_A \hat{H} \psi_A \, d\tau = \int \psi_B \hat{H} \psi_B \, d\tau$ (essentially the original atomic-orbital energy, perturbed by the presence of the other nucleus), and the **resonance integral** $\beta = \int \psi_A \hat{H} \psi_B \, d\tau$ (negative, and nonzero only because the orbitals overlap). Normalizing $\psi_{\mathrm{MO}}^{\pm} = (\psi_A \pm \psi_B)/\sqrt{2(1 \pm S)}$ and evaluating $\langle \psi_{\mathrm{MO}}^{\pm} | \hat{H} | \psi_{\mathrm{MO}}^{\pm} \rangle$ gives

$$E_{\pm} = \frac{\alpha \pm \beta}{1 \pm S}. \tag{12.2}$$

Because $\beta < 0$, and because $|\beta|$ is large enough compared to $|\alpha| S$ for essentially every real covalent bond (the case of interest here), $E_{+} = (\alpha + \beta)/(1 + S)$ lies below α (bonding) and $E_{-} = (\alpha - \beta)/(1 - S)$ lies above α (antibonding) — but the two shifts are **not equal in magnitude**. The $(1 - S)$ in the antibonding denominator is smaller than the $(1 + S)$ in the

bonding denominator, so E_- rises above α by more than E_+ falls below it. This asymmetry is the entire reason He_2 fails to bond: two electrons in σ_{1s} and two in σ^*_{1s} do not cancel to zero net energy change, because the antibonding pair's destabilization outweighs the bonding pair's stabilization, leaving He_2 at *higher* energy than two separate helium atoms — consistent with, and quantitatively explaining, the bond order of zero found by simple electron counting below.

Filling the resulting molecular orbitals with the molecule's electrons, two at a time (spin-paired, per the exclusion principle applied now to molecular rather than atomic orbitals) from lowest to highest energy, gives a **molecular orbital diagram**, from which the **bond order** is computed as

$$\text{bond order} = \frac{(\text{number of bonding electrons}) - (\text{number of antibonding electrons})}{2}. \quad (12.3)$$

Bond order need not be an integer. Whenever an odd number of electrons occupies an antibonding level, the formula returns a half-integer, such as bond order $\frac{1}{2}$ for He_2^+ or $\frac{5}{2}$ for N_2^+ — a genuine prediction of MO theory, not a rounding artifact.

A bond order of zero predicts no net covalent energy lowering relative to separated atoms; a bond order of $1, 2, 3, \ldots$ corresponds roughly to a single, double, triple, ... bond, with higher bond order generally correlating with a shorter, stronger bond. For H_2 (two electrons, both in σ_{1s}), the bond order is $(2-0)/2 = 1$, consistent with the known stable single bond; for the hypothetical He_2 (four electrons, two in σ_{1s} and two forced by the exclusion principle into σ^*_{1s}), the bond order is $(2-2)/2 = 0$ — correctly predicting that He_2 has no conventional covalent bond, a conclusion valence bond theory (which has no natural way to place electrons in an antibonding orbital) does not straightforwardly reach. A neutral helium dimer can nevertheless exist as an extraordinarily weak van der Waals-bound state, an intermolecular effect outside this simple covalent MO diagram. MO theory additionally predicts a molecule's magnetic behavior directly from its orbital diagram: any unpaired electrons (occurring, per Hund's rule applied to degenerate molecular orbitals, when a set of same-energy orbitals is only partially filled) make the molecule **paramagnetic** (weakly attracted into a magnetic field), while a fully paired configuration makes it **diamagnetic** (weakly repelled) — famously correctly predicting that O_2 is paramagnetic (two unpaired electrons in degenerate antibonding π^* orbitals), a fact simple Lewis-structure/valence-bond reasoning does not anticipate.

Antibonding Is Not the Same as Nonbonding

It is tempting to think of an antibonding orbital as simply "not contributing" to the bond, the way a lone pair sitting in an uninvolved orbital does. That is wrong: an antibonding orbital actively *raises* the energy of an electron placed in it above the energy of the separated atomic orbitals, because destructive interference at the internuclear midpoint removes electron density from between the nuclei rather than merely failing to add any. A fully occupied antibonding orbital does not leave a bond unaffected — paired with a corresponding fully occupied bonding orbital, it cancels the bond outright, exactly as for He_2 above. "Antibonding" means actively destabilizing, not neutral.

The bonding and antibonding pair are not a chemical convention; they are what the Schrödinger equation returns for two wells brought close together, and Figure 12.3 solves

that problem directly. Start with two widely separated square wells: each has its own ground state, and the two are degenerate. Slide them together and the degeneracy lifts into exactly two states — one symmetric, with no node between the wells and an energy *below* the isolated-atom level, and one antisymmetric, with a node at the midpoint and an energy above it. That is σ_{1s} and σ^*_{1s}, obtained without mentioning chemistry, and the splitting between them grows as the wells approach, which is why bond strength depends on overlap.

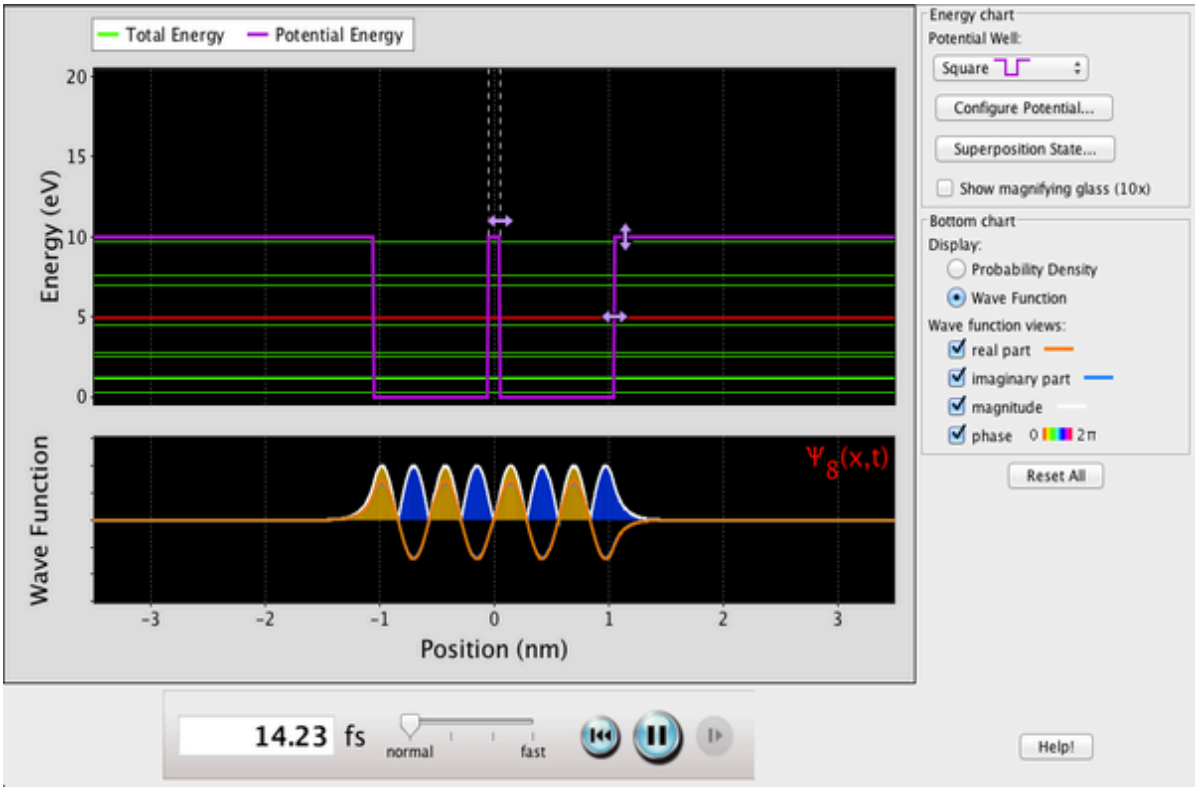

Figure 12.3: A double square well with adjustable separation and depth, and the eigenstates it supports. The symmetric–antisymmetric splitting of a pair of formerly degenerate levels is the origin of the bonding/antibonding pair of MO theory, and — extended to N wells in a row — of the energy bands of a solid. Interactive simulation: Double Wells and Covalent Bonds

Historical Context: From Lewis's Electron Pair to the Pauling–Mulliken Rivalry

The idea that a covalent bond is a *shared pair of electrons* predates quantum mechanics itself. In 1916 — a full decade before Schrödinger's equation — the American chemist Gilbert N. Lewis proposed exactly this picture, together with the electron-dot notation still used to sketch molecules today, purely from chemical reasoning about valence and the stability of the noble-gas electron count, with no wave mechanics available to justify it. Lewis's shared pair was a hypothesis in search of a mechanism, and quantum mechanics supplied one a decade later, in two competing forms.

Valence bond theory, developed by Walter Heitler and Fritz London (1927) and then extended and popularized by Linus Pauling through the early 1930s — including the hybrid-orbital concept of this chapter — kept Lewis's picture of a bond as a localized pair shared between two specific atoms, now computed from overlapping atomic wave functions rather than guessed at. **Molecular orbital theory**, developed in the same years chiefly by Friedrich

Hund and Robert Mulliken, instead discarded the idea that an electron belongs to one bond at a time, building orbitals delocalized over the whole molecule from the outset. The two camps disagreed, sometimes sharply, over which picture was the physically correct starting point — Pauling's localized, chemist-friendly bonds versus Mulliken's delocalized, spectroscopically motivated orbitals — and valence bond theory, propelled by Pauling's enormously influential 1939 book *The Nature of the Chemical Bond*, remained the dominant teaching framework through the 1950s. Molecular orbital theory eventually overtook it, in no small part on the strength of results like the O_2 paramagnetism prediction below, which valence bond theory cannot produce without ad hoc patching. Pauling received the 1954 Nobel Prize in Chemistry "for his research into the nature of the chemical bond," and Mulliken received the 1966 Prize "for his fundamental work concerning chemical bonds and the electronic structure of molecules by the molecular orbital method" (Figure 12.4). Both theories remain in active use today, each suited to different questions — geometry and localized reactivity for valence bond theory, spectra and magnetism for molecular orbital theory — which is why this chapter, like the field itself, teaches both.

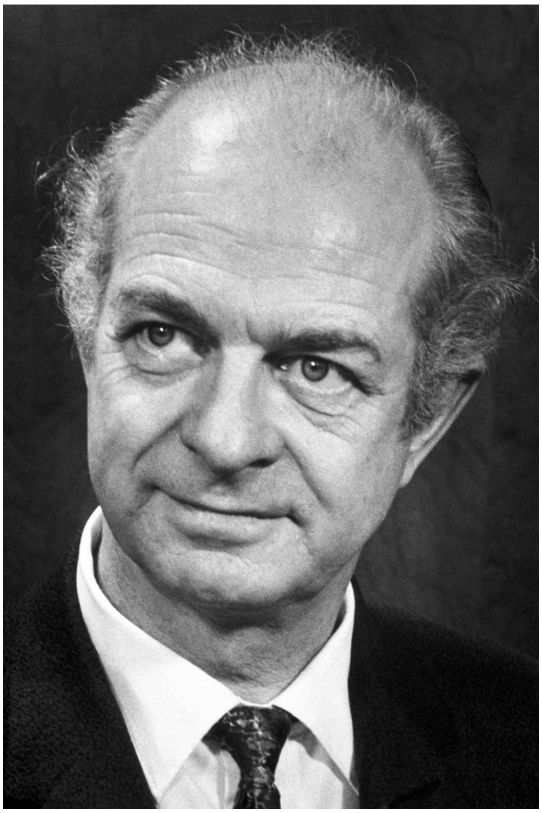

Figure 12.4: Linus Pauling, photographed in 1962, eight years after his 1954 Nobel Prize in Chemistry for work on the nature of the chemical bond — the valence-bond and hybridization framework of this chapter. Photograph by the Nobel Foundation; public domain via Wikimedia Commons.

Building the Diagram for the Second Row: N_2 and O_2

Hydrogen and helium have only 1*s* orbitals to combine, but a second-row diatomic such as N_2 or O_2 has both 2*s* and 2*p* atomic orbitals on each atom, and the resulting molecular orbital diagram, while built by exactly the same LCAO recipe used for H_2, has more structure worth

spelling out explicitly.

Combining the two $2s$ orbitals (one per atom) gives a bonding σ_{2s} and an antibonding σ^*_{2s}, exactly as for $1s$. Of the three $2p$ orbitals on each atom, the pair pointing directly *along* the internuclear axis overlaps head-on, giving a bonding σ_{2p} and antibonding σ^*_{2p}; the two remaining pairs of $2p$ orbitals, oriented *perpendicular* to the axis, overlap sideways, giving a bonding π_{2p} and antibonding π^*_{2p} pair for each of the two independent perpendicular directions — so the π_{2p} level (and, separately, the π^*_{2p} level) is **doubly degenerate**, two orbitals at the same energy. In all, eight atomic orbitals (one $2s$ and three $2p$ on each of two atoms) combine into eight molecular orbitals: σ_{2s}, σ^*_{2s}, σ_{2p}, σ^*_{2p}, and the degenerate pairs π_{2p} and π^*_{2p}.

The one genuine subtlety is the *relative* energy of σ_{2p} and π_{2p}, and it is worth getting right because it is a famous trap. Head-on (σ) overlap is intrinsically stronger than sideways (π) overlap, so a naive picture would put σ_{2p} below π_{2p} always — and that is indeed the order for O_2, F_2, and Ne_2. But for the *lighter* diatomics B_2, C_2, and N_2, the order **inverts**: π_{2p} sits below σ_{2p}. The mechanism is s–p **mixing**: σ_{2s} and σ_{2p} share the same symmetry along the internuclear axis, and whenever the atomic $2s$ and $2p$ energies are close enough, these two molecular orbitals mix with each other quantum mechanically, pushing σ_{2s} lower and σ_{2p} higher than a naive non-interacting picture would predict — enough, for the lighter elements, to push σ_{2p} above π_{2p} entirely. Moving across the period, increasing nuclear charge pulls the $2s$ orbital down in energy faster than the $2p$ orbital (a $2s$ electron spends more time near the nucleus and so feels the increasing charge more strongly), widening the $2s$–$2p$ gap; by oxygen, that gap is wide enough that s–p mixing is too weak to invert the order, and the naive σ_{2p}-below-π_{2p} sequence is restored. Figure 12.5 shows both diagrams side by side, filled with electrons.

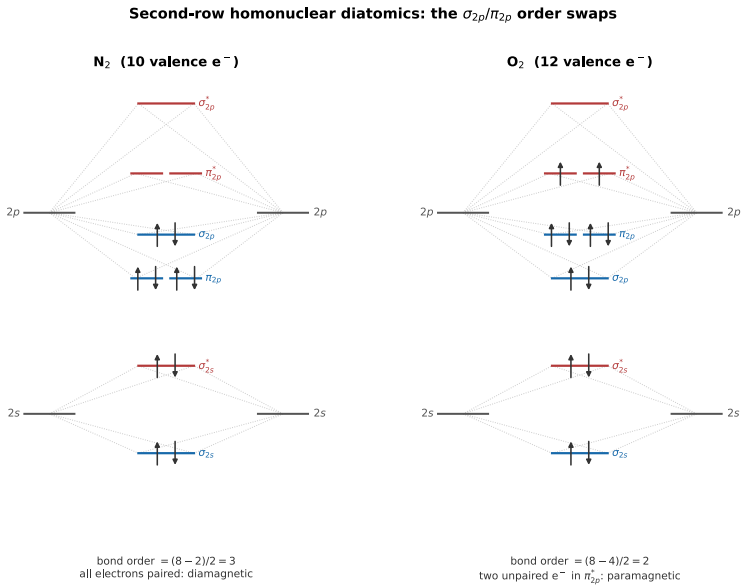

Figure 12.5: Molecular orbital diagrams for N_2 and O_2. Atomic $2s$ and $2p$ levels sit outside; molecular orbitals sit in the middle, filled from the bottom with the molecule's valence electrons. Note the swapped σ_{2p}/π_{2p} order: π_{2p} below σ_{2p} for N_2 (light diatomic, strong s–p mixing), σ_{2p} below π_{2p} for O_2 (heavier, weak s–p mixing). Original schematic generated with matplotlib; see `scripts/figures/`.

For N_2 (10 valence electrons), filling from the bottom — σ_{2s} (2), σ^*_{2s} (2), π_{2p} (4), σ_{2p} (2) — uses all ten electrons with every orbital either completely filled or completely empty, giving

Table 12.2: Filling the O2 molecular orbital diagram

Orbital	Electrons added	Running total
σ_{2s}	2	2
σ^*_{2s}	2	4
σ_{2p}	2	6
π_{2p} (both orbitals)	4	10
π^*_{2p} (both orbitals)	2	12

$$\text{bond order}(N_2) = \frac{8-2}{2} = 3, \tag{12.4}$$

consistent with the triple bond of the Lewis structure : N ≡ N :, and, with every electron paired, correctly predicting that N_2 is diamagnetic.

Worked Example: The Molecular Orbital Diagram and Paramagnetism of O_2

Construct the molecular orbital diagram for O_2, determine its bond order, and explain its paramagnetism.

O_2 has $6+6=12$ valence electrons. Using the *heavier-diatomic* ordering established above — σ_{2s}, σ^*_{2s}, σ_{2p}, π_{2p} (×2), π^*_{2p} (×2), σ^*_{2p} — fill from the bottom, two electrons per orbital, respecting Hund's rule (spread electrons across a set of degenerate orbitals, one each, before pairing any of them), as in Table 12.2:

The last two electrons enter the doubly degenerate π^*_{2p} level. Hund's rule places one electron in each of the two degenerate π^*_{2p} orbitals, with parallel spins, rather than pairing both into a single orbital — leaving **two unpaired electrons**. Counting bonding electrons (σ_{2s}: 2, σ_{2p}: 2, π_{2p}: 4, total 8) against antibonding electrons (σ^*_{2s}: 2, π^*_{2p}: 2, total 4),

$$\text{bond order}(O_2) = \frac{8-4}{2} = 2, \tag{12.5}$$

consistent with the double bond of the Lewis structure O = O. But that Lewis structure, with every electron paired off into bonds and lone pairs, gives no hint of the two unpaired π^*_{2p} electrons found here — and it is exactly those two unpaired electrons that make liquid oxygen **paramagnetic**, visibly drawn toward the poles of a strong magnet in the standard classroom demonstration, a fact valence bond theory cannot explain without modification but that falls directly out of the molecular orbital diagram.

Bond Order, Bond Length, and Bond Strength: A Worked Comparison

Bond order is a prediction about two directly measurable quantities: bond length and bond dissociation energy. Higher bond order means more shared electron density concentrated between the nuclei, which pulls the nuclei closer together (shorter bond length) and requires more energy to pull them apart (higher dissociation energy). The nitrogen–nitrogen bond, compared across three different molecules in Table 12.3, shows the trend cleanly, because in each case the bond order is unambiguous from the Lewis structure:

Tripling the bond order roughly quintuples the dissociation energy while shortening the bond by nearly a quarter — and the relationship is not linear in dissociation energy: going from a single to a double bond adds 251 kJ/mol, while going from a double to a triple bond

Table 12.3: Nitrogen–nitrogen bond order, length, and dissociation energy

Molecule	Bond	Bond order	Bond length	Dissociation energy
N_2H_4 (hydrazine)	N–N	1	145 pm	167 kJ/mol
N_2H_2 (diazene)	N=N	2	125 pm	418 kJ/mol
N_2 (nitrogen)	N≡N	3	110 pm	942 kJ/mol

adds 524 kJ/mol, more than double the first increment, because the additional π bonds of a multiple bond form between orbitals already held close together by the existing σ bond and so overlap unusually well. This is also why N_2, held together by one of the strongest common bonds in chemistry, is so notoriously unreactive that converting it into a chemically usable form of nitrogen (the industrial Haber–Bosch process) is one of the most energy-intensive reactions carried out on Earth.

Read the HOMO Before You Refill the Whole Diagram

When a problem asks how ionizing (or reducing) a molecule changes its bond length or bond strength, there is rarely a need to refill an entire molecular orbital diagram from scratch. Identify only the **highest occupied molecular orbital (HOMO)** — the orbital the added or removed electron actually occupies — and ask whether it is bonding or antibonding. Removing an electron from a bonding HOMO weakens the bond (as for $N_2 \rightarrow N_2^+$ below); removing one from an antibonding HOMO strengthens it (as for $O_2 \rightarrow O_2^+$ in Exercise 12.6). That one fact settles the direction of the change before a bond order is ever computed.

Worked Example: Ionizing Nitrogen — N_2 versus N_2^+

Removing an electron from N_2 to form the molecular ion N_2^+ (as happens, for example, when a fast electron or solar-wind particle strikes an atmospheric nitrogen molecule during an aurora) removes it from the highest-occupied orbital identified above — for N_2, the bonding σ_{2p}. That leaves 9 valence electrons: σ_{2s} (2), σ^*_{2s} (2), π_{2p} (4), σ_{2p} (1), so

$$\text{bond order}(N_2^+) = \frac{7-2}{2} = 2.5, \tag{12.6}$$

down from 3 for neutral N_2. Removing an electron from a *bonding* orbital weakens the bond, predicting a longer, weaker bond in the ion than in the neutral molecule — and indeed N_2^+'s measured bond length, about 112 pm, is longer than N_2's 110 pm, consistent with its lower bond order. The single unpaired electron left in σ_{2p} also makes N_2^+ paramagnetic, in contrast to diamagnetic neutral N_2; the characteristic blue emission of the aurora's nitrogen-ion band is, in fact, this very ion relaxing from an excited electronic state.

Compare this with the O_2 family of Exercise 12.6: there, removing an electron from the *antibonding* π^*_{2p} orbital *strengthens* the bond rather than weakening it, because O_2's highest-occupied orbital is antibonding while N_2's is bonding. The same operation — ionization, removing one electron — can strengthen or weaken a bond depending entirely on the character of the specific orbital the electron is removed from, information only the molecular orbital diagram supplies and that a Lewis structure alone cannot.

12.3 Intermolecular Forces and Molecular Spectra

Intermolecular Forces

Everything so far in this chapter describes **intramolecular** forces — the bonds that hold the atoms of a single molecule together, whether ionic, covalent, or described by a molecular orbital diagram. Bulk matter — liquids and solids made of many molecules — depends just as much on much weaker **intermolecular forces**, the attractions *between* separate, already-bonded molecules. These forces are typically one to two orders of magnitude weaker than a covalent bond (tens of kJ/mol, rather than hundreds), but they are exactly what must be overcome to melt a solid or boil a liquid, and their strength — not the strength of the covalent bonds within each molecule, which survive melting and boiling completely intact — is what actually sets a substance's melting and boiling points.

London dispersion forces act between *every* pair of molecules, regardless of polarity. A molecule's electron cloud fluctuates from instant to instant, producing a fleeting, temporary dipole moment even in a molecule with no permanent dipole at all; that instantaneous dipole induces a matching temporary dipole in a neighboring molecule, and the two weakly attract. Dispersion forces strengthen with a molecule's **polarizability** — how easily its electron cloud is distorted — which in turn grows with the number of electrons and the physical size of the molecule, which is why boiling points climb steadily up a family of increasingly large nonpolar molecules (the noble gases, or the halogens) even though none of them has a permanent dipole moment at all.

Dipole–dipole forces act between molecules that already carry a *permanent* dipole moment (recall Figure 12.1, above): the positive end of one polar molecule is attracted to the negative end of its neighbor. Because this attraction does not have to wait on a random fluctuation, it is generally stronger, molecule for molecule, than a dispersion-only attraction between molecules of comparable size.

Hydrogen bonding is an unusually strong special case of dipole–dipole attraction, occurring specifically when a hydrogen atom is bonded directly to a small, highly electronegative atom — nitrogen, oxygen, or fluorine — leaving that hydrogen with a large, concentrated partial positive charge (it has essentially no core electrons of its own to shield its bare proton) that is then strongly attracted to a lone pair on an N, O, or F atom of a neighboring molecule. Hydrogen bonds are typically five to ten times stronger than an ordinary dipole–dipole attraction, though still far weaker than a covalent bond, and are responsible for water's unusually high boiling point, the double-helix structure of DNA (hydrogen bonds between complementary base pairs), and the open crystal structure that makes ice less dense than the liquid water it floats on.

Worked Example: Dispersion versus Dipole–Dipole — Butane and Acetone

Butane (C_4H_{10}, nonpolar, molar mass 58.1 g/mol) and acetone (C_3H_6O, polar, molar mass 58.1 g/mol) have essentially identical molar mass — and therefore comparable numbers of electrons, comparable polarizability, and hence comparable London dispersion forces — yet butane boils at -1°C while acetone boils at 56°C, a difference of 57 Celsius degrees. Since dispersion forces are approximately equal for the two molecules, most of the difference comes from an attraction dispersion forces alone cannot supply: acetone's carbonyl group (C=O) carries a substantial permanent dipole moment (about 2.9 D), giving it dipole–dipole attractions that butane, with no permanent dipole at all, simply lacks. Molecular shape also affects how

efficiently molecules pack and interact, so matching molecular size isolates dipole–dipole attraction as the main additional factor rather than the sole possible source of the boiling-point difference.

The same logic, pushed one step further, isolates hydrogen bonding specifically. Water (H_2O, molar mass 18 g/mol) boils at 100°C, while hydrogen sulfide (H_2S, molar mass 34 g/mol, the next member down the same column of the periodic table) boils at −60°C — even though the *heavier*, more polarizable H_2S should, by dispersion forces alone, boil *higher* than water, not lower. Water's anomalously high boiling point is the signature of hydrogen bonding: each water molecule can form up to four hydrogen bonds (two through its own hydrogens, two through its oxygen's lone pairs), building an extended three-dimensional network that an ordinary dipole–dipole liquid like H_2S — whose S–H bond is too weakly polar for effective hydrogen bonding — never forms.

Vibrational and Rotational Energy Levels

Why Electrons and Nuclei Can Be Treated Separately: The Born–Oppenheimer Approximation

Every molecular orbital diagram and hybridization argument earlier in this chapter implicitly treats the nuclei as fixed in place while the electrons rearrange around them, and the analysis that follows now treats the nuclei as vibrating and rotating about a bond length r_0 taken as already known, with no further reference to the electrons at all. This split is justified by the **Born–Oppenheimer approximation**, proposed by Max Born and J. Robert Oppenheimer in 1927: because a nucleus is thousands of times more massive than an electron (a lone proton alone outweighs an electron by a factor of about 1836), the electrons respond to any change in nuclear position essentially instantaneously on the timescale of nuclear motion. The electronic structure problem can therefore be solved first, at each fixed internuclear separation, to obtain an effective potential energy curve for the nuclei — precisely the curve plotted in Figure 12.6 below — after which the vibration and rotation of the nuclei on that curve become a separate, subsequent problem. Without this approximation, the full molecular Schrödinger equation, coupling every electron's motion to every nucleus's motion at once, would be intractable for all but the smallest molecules.

Once bonded, a diatomic molecule is itself a quantum system with its own internal energy levels, in addition to the electronic energy levels associated with its bonding orbitals. Near the equilibrium bond length r_0 (where the molecular potential energy curve, as a function of internuclear separation, has its minimum), the potential is well approximated by a parabola, so small-amplitude **vibration** of the two nuclei about r_0 is, to good approximation, the quantum harmonic oscillator of Chapter 8, with quantized energies

$$E_v = \left(v + \tfrac{1}{2}\right)\hbar\omega, \qquad v = 0, 1, 2, \ldots, \tag{12.7}$$

where $\omega = \sqrt{k/\mu}$, k is the effective "spring constant" of the bond (obtained from the curvature of the potential at its minimum), and $\mu = m_1 m_2/(m_1 + m_2)$ is the **reduced mass** of the two-nucleus system (the appropriate effective mass for relative motion of a two-body system, reducing the two-body vibration problem to an equivalent single-particle problem).

The parabola is an approximation, and it is useful to see what it approximates. Figure 12.6 plots the interatomic potential energy of a diatomic pair as a function of separation: strongly

repulsive at short range, attractive at long range, with a minimum at r_0 whose depth is the bond energy and whose curvature is the k in $\omega = \sqrt{k/\mu}$. Pull the atoms far from r_0 and the curve is visibly not a parabola — it flattens out toward dissociation on one side and rises much faster than quadratically on the other — which is why real molecular vibrational levels crowd together at high v instead of staying evenly spaced at $\hbar\omega$.

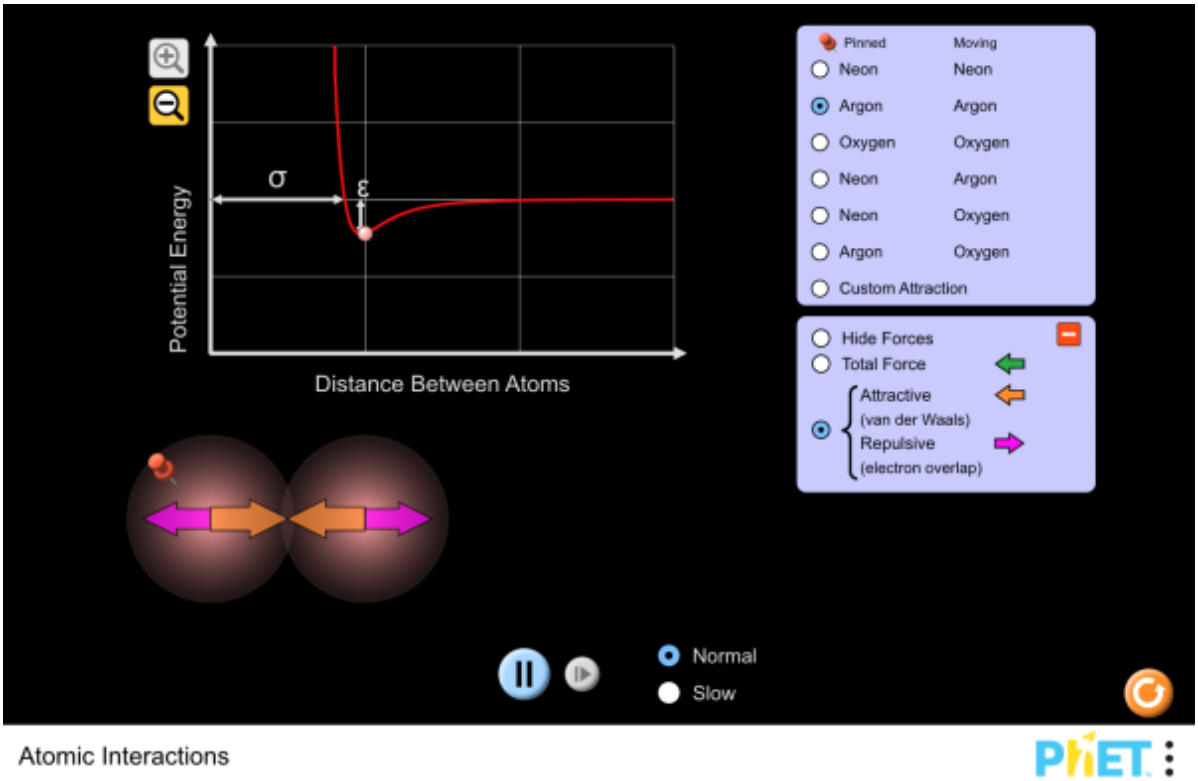

Figure 12.6: The potential energy curve between two atoms, with the atom types adjustable. The equilibrium separation, the well depth, and the curvature at the minimum are the bond length, the bond energy, and the vibrational spring constant respectively. Interactive simulation: Atomic Interactions

Independently, the molecule can **rotate** about its center of mass; treating the two nuclei as point masses at fixed separation r_0 (the **rigid rotor** approximation, reasonable when rotational energies are small compared to vibrational spacing) makes this exactly the angular-momentum problem of Chapter 9, with quantized rotational energy

$$E_J = \frac{\hbar^2}{2I} J(J+1), \qquad J = 0, 1, 2, \ldots, \tag{12.8}$$

where $I = \mu r_0^2$ is the molecule's moment of inertia and J plays the role of the orbital angular momentum quantum number ℓ. Because I for a typical molecule is large (bond lengths of order 10^{-10} m, but heavy nuclear masses) compared to the effective "moment of inertia" scale set by an electron, rotational energy spacings are much smaller than vibrational spacings, which are in turn much smaller than electronic transition energies — a hierarchy ($E_{\text{elec}} \gg E_{\text{vib}} \gg E_{\text{rot}}$) that is directly reflected in molecular spectra: electronic transitions lie in the visible/ultraviolet, vibrational transitions in the infrared, and pure rotational transitions in the microwave region, each region probing a different aspect of molecular structure.

This hierarchy is not merely a table of numbers; it is why a molecule responds to one part

of the spectrum and ignores another, and Figure 12.7 lets it be tested one photon at a time. Aim microwaves at a molecule and it rotates. Switch to infrared and it starts to vibrate — but only if the vibration changes the molecule's dipole moment, which is why N_2 and O_2 have no strong fundamental electric-dipole absorption in the infrared while CO_2 and H_2O do, and hence why the two minor constituents of the atmosphere, not the two major ones, set the temperature of the planet. At sufficiently short ultraviolet wavelengths, a photon can excite an electronic state and may have enough energy to dissociate the molecule.

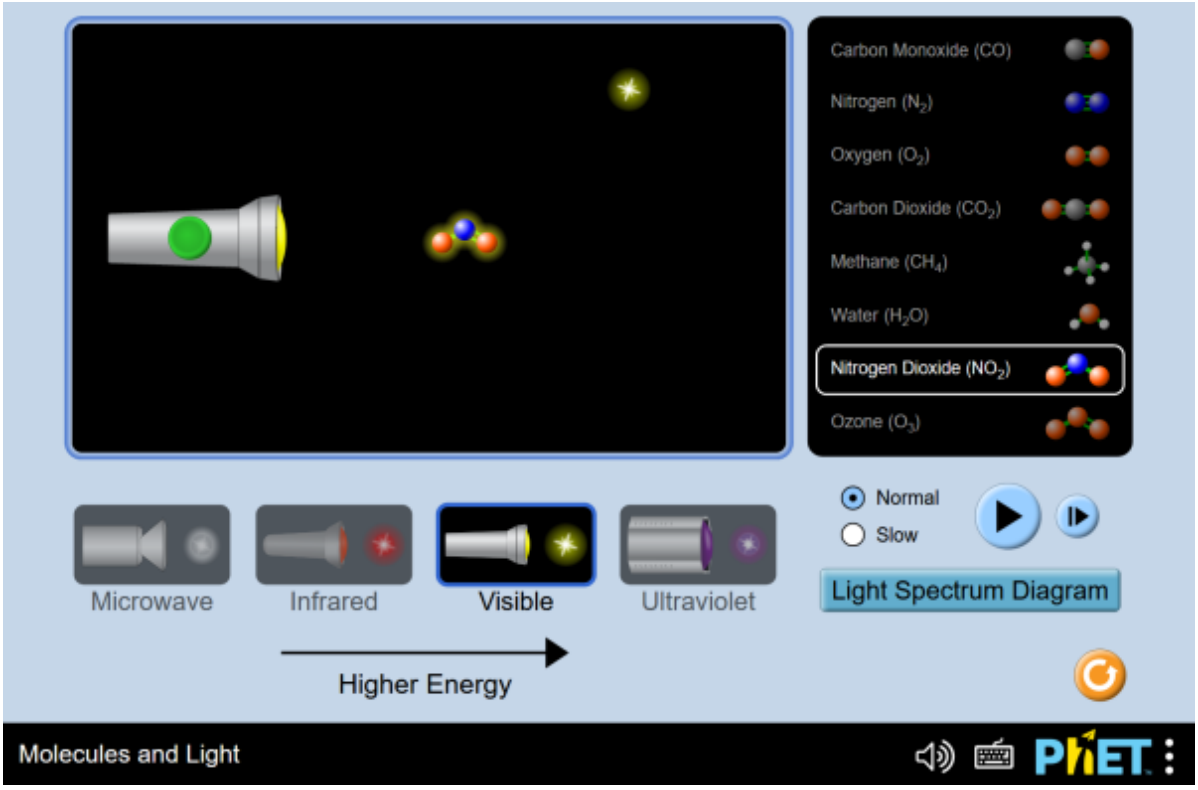

Figure 12.7: Single photons of a chosen wavelength directed at a chosen molecule. Microwave, infrared, visible, and ultraviolet photons each produce a different response — rotation, vibration, electronic excitation, or dissociation — in the order of the $E_{\text{elec}} \gg E_{\text{vib}} \gg E_{\text{rot}}$ hierarchy. Interactive simulation: Molecules and Light

The Rovibrational Spectrum: Vibration and Rotation Together

A real infrared absorption spectrum does not show the vibrational transition as a single line at $\hbar\omega$. Because a molecule is simultaneously vibrating *and* rotating, a photon absorbed in a vibrational transition ($\Delta v = +1$) is generally accompanied by a simultaneous change in rotational state, subject to the selection rule

$$\Delta J = \pm 1 \tag{12.9}$$

(for a diatomic molecule, whose single vibrational mode necessarily oscillates the dipole moment along the bond axis itself, quantum-mechanical selection rules for this type of vibration forbid $\Delta J = 0$ — the analog of a missing Q branch — leaving only $\Delta J = \pm 1$ available; some polyatomic vibrations that shift the dipole moment perpendicular to a molecular symmetry

axis do permit a weak Q branch, but a diatomic never has that option). Writing the combined vibration–rotation energy as $E_{v,J} = \left(v + \frac{1}{2}\right)\hbar\omega + BJ(J+1)$, with the **rotational constant** $B \equiv \hbar^2/2I$ (an energy, in the convention used throughout this chapter), the photon energy absorbed in a transition from $(v = 0, J)$ to $(v = 1, J')$ is

$$h\nu = \hbar\omega + B\big[J'(J'+1) - J(J+1)\big]. \tag{12.10}$$

Two families of lines result, sketched in Figure 12.8. The **R branch** ($\Delta J = +1$, $J' = J+1$) works out to $h\nu = \hbar\omega + 2B(J+1)$ for $J = 0, 1, 2, \ldots$, giving lines above $\hbar\omega$ spaced by $2B$. The **P branch** ($\Delta J = -1$, $J' = J - 1$) works out to $h\nu = \hbar\omega - 2BJ$ for $J = 1, 2, 3, \ldots$, giving lines below $\hbar\omega$, also spaced by $2B$. No line appears at $h\nu = \hbar\omega$ itself — that would be the forbidden $\Delta J = 0$ **Q branch** — leaving a characteristic gap of about $4B$ at the center of the band, a gap that is itself a direct, measurable signature of the selection rule.

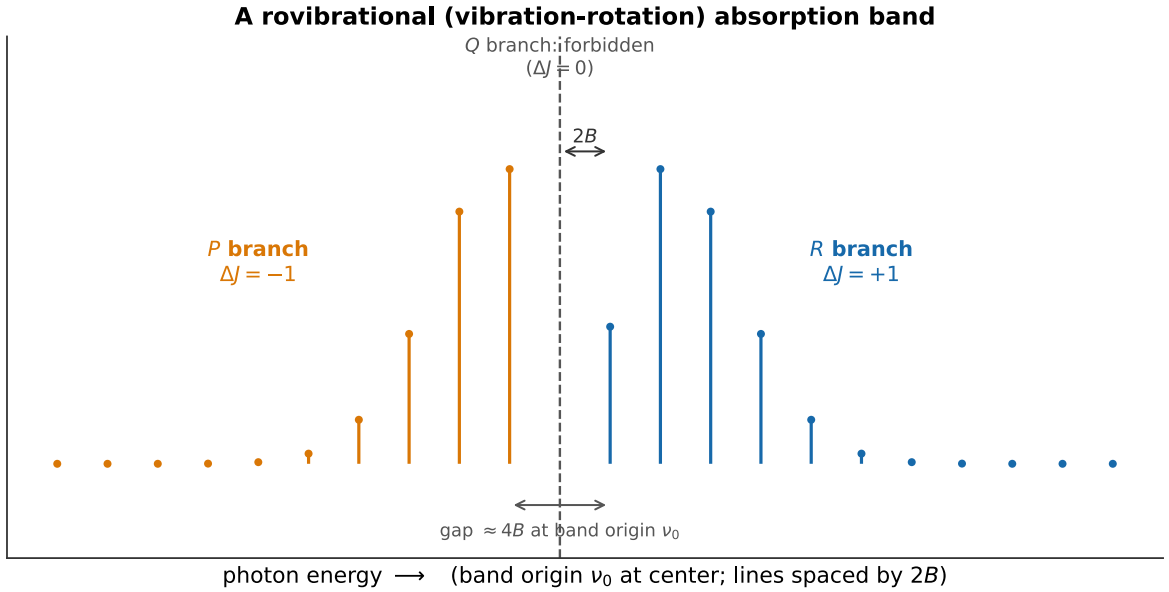

Figure 12.8: A rovibrational absorption band. The R branch ($\Delta J = +1$) and P branch ($\Delta J = -1$) each consist of lines spaced by $2B$; the missing Q branch ($\Delta J = 0$, forbidden for a diatomic) leaves a gap of about $4B$ at the band origin. Line intensities (schematic here) track the thermal population of each rotational level before absorption.

Because B depends only on the molecule's moment of inertia, measuring the line spacing in a single rovibrational (infrared) spectrum determines I — and hence the bond length r_0 — directly, without needing a separate microwave (pure-rotation) measurement at all.

> A **wavenumber**, $\tilde{\nu} \equiv E/hc$, reports energy in units of cm^{-1} — the number of wave cycles per centimeter for a photon of that energy. It is the conventional unit throughout infrared and microwave spectroscopy, chosen because a spectrometer directly measures wavelength or frequency, not joules.

Worked Example: Line Spacing in the CO Rovibrational Spectrum

Using the same carbon monoxide data as Exercise 12.5 ($r_0 = 0.113$ nm, $\mu = 6.86$ u), find the rotational constant B and the resulting rovibrational line spacing.

The moment of inertia is

$$I = \mu r_0^2 = (6.86)(1.66 \times 10^{-27}\ \text{kg})(0.113 \times 10^{-9}\ \text{m})^2 = 1.45 \times 10^{-46}\ \text{kg} \cdot \text{m}^2, \tag{12.11}$$

so

$$B = \frac{\hbar^2}{2I} = \frac{(1.055 \times 10^{-34}\ \text{J} \cdot \text{s})^2}{2(1.45 \times 10^{-46}\ \text{kg} \cdot \text{m}^2)} = 3.83 \times 10^{-23}\ \text{J} = 2.39 \times 10^{-4}\ \text{eV}. \tag{12.12}$$

Both branches are spaced by $2B = 4.78 \times 10^{-4}$ eV — the same energy as the $J = 0 \to J = 1$ pure rotational transition of Exercise 12.5, as it must be, since both quantities are just $2B$ measured two different ways. Converting to the wavenumber units ($\tilde{\nu} \equiv E/hc$) conventional in infrared spectroscopy,

$$\tilde{B} = \frac{B}{hc} = \frac{3.83 \times 10^{-23}\ \text{J}}{(6.626 \times 10^{-34}\ \text{J} \cdot \text{s})(2.998 \times 10^{10}\ \text{cm/s})} = 1.93\ \text{cm}^{-1}, \tag{12.13}$$

so the predicted line spacing is $2\tilde{B} = 3.85\ \text{cm}^{-1}$ — in excellent agreement with the spacing of about $3.86\ \text{cm}^{-1}$ actually observed in the CO fundamental infrared band. A bond length and a reduced mass, fed into a formula derived from nothing more than the rigid-rotor approximation, correctly predict the fine structure of a real molecular spectrum.

Anharmonicity Revisited

Figure 12.6, above, already showed that the true interatomic potential is not a perfect parabola: it rises more steeply than quadratic at short range (the repulsive wall) and flattens out well below quadratic at long range, approaching the dissociation energy asymptotically rather than climbing forever. A more realistic potential — the **Morse potential** is the standard choice for modeling this curve quantitatively — gives vibrational energy levels that are no longer *exactly* evenly spaced:

$$E_v \approx \left(v + \tfrac{1}{2}\right)\hbar\omega - \left(v + \tfrac{1}{2}\right)^2 x_e \hbar\omega, \qquad v = 0, 1, 2, \ldots, \tag{12.14}$$

where the small, positive **anharmonicity constant** x_e (typically a few percent, for real molecules) quantifies the departure from a perfect harmonic oscillator. The negative correction term grows with v, so **the levels crowd closer together at higher** v — an effect invisible near the bottom of the well, where the potential is well approximated by the parabola of the harmonic-oscillator treatment above, but increasingly important as v climbs toward the dissociation limit, where the level spacing approaches zero as the bound levels approach the molecule's bond dissociation energy. Anharmonicity also relaxes the strict $\Delta v = \pm 1$ selection rule of the ideal harmonic oscillator, permitting weak **overtone** transitions with $\Delta v = \pm 2, \pm 3, \ldots$ — additional, much fainter absorption lines at roughly (but not exactly) integer multiples of the fundamental frequency. This is why a real vibrational spectrum shows one strong fundamental band accompanied by a series of progressively weaker overtones, rather than the single, perfectly sharp line the ideal harmonic oscillator of Chapter 8 would predict.

A Bound System Weighs Less Than Its Parts, Again
A bond's dissociation energy — the energy needed to pull two bonded atoms apart to infinite separation — is the molecular-scale instance of a pattern that reappears, at a vastly different energy and length scale, when the same nuclei are pulled apart from *within*. See Nuclear Physics for the nuclear binding energy, where the same logic (a bound system has lower total energy, and hence lower total mass, than its separated constituents) explains why nuclei resist being split or fused apart, on a scale of MeV rather than kJ/mol.

12.4 Summary

- Chemical bonds lower a molecule's total energy relative to separated atoms; **ionic bonding** (electron transfer, electrostatic attraction) and **covalent bonding** (shared electron pairs, enhanced internuclear electron density) are limiting cases of a continuum set by electronegativity difference.
- **Valence bond theory** builds bonds from overlapping atomic orbitals (σ for direct, π for sideways overlap); **hybrid orbitals** (sp, sp^2, sp^3, etc.), one set per number of electron domains, reproduce observed molecular geometries via minimization of electron-pair repulsion.
- **Molecular orbital theory** builds orbitals belonging to the whole molecule as linear combinations of atomic orbitals, splitting into lower-energy **bonding** and higher-energy **antibonding** orbitals; filling these with the molecule's electrons gives the **bond order**, which predicts stability, bond strength, and (via unpaired electrons) paramagnetism — correctly predicting, e.g., O_2's paramagnetism and the absence of a conventional covalent bond in He_2.
- A bonded diatomic molecule has quantized **vibrational** levels, $E_v = (v+\frac{1}{2})\hbar\omega$ (harmonic oscillator in the reduced mass μ), and **rotational** levels, $E_J = \hbar^2 J(J+1)/2I$ (rigid rotor), with $E_{\text{elec}} \gg E_{\text{vib}} \gg E_{\text{rot}}$, placing electronic, vibrational, and rotational spectra in the UV/visible, infrared, and microwave regions respectively.
- The σ_{2p}/π_{2p} **ordering inverts** between light second-row diatomics (B_2–N_2, π_{2p} lower via strong s–p mixing) and heavier ones (O_2–Ne_2, σ_{2p} lower); higher bond order predicts a shorter, stronger bond, confirmed across N–N single/double/triple bonds and by ionizing N_2, O_2, and F_2.
- **Intermolecular forces** — London dispersion (universal, grows with polarizability), dipole–dipole (permanent dipoles), and **hydrogen bonding** (an unusually strong dipole–dipole interaction specific to H bonded to N, O, or F) — act *between* already-bonded molecules and set melting and boiling points; comparing molecules of similar size isolates each contribution.
- A **rovibrational spectrum** combines a vibrational transition with a simultaneous rotational one, producing R-branch ($\Delta J = +1$) and P-branch ($\Delta J = -1$) lines spaced by $2B = \hbar^2/I$, with a missing central Q branch ($\Delta J = 0$, forbidden for a diatomic); **anharmonicity** of the real interatomic potential crowds vibrational levels together at high v and permits weak overtone transitions.

12.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens on molecular-orbital diagrams for nitrogen and oxygen.

1. **Multiple choice, from a figure.** Why is molecular oxygen paramagnetic in the molecular-orbital description? Figure description: Molecular-orbital energy-level diagrams for nitrogen and oxygen. Choices: (A) It has two unpaired electrons in antibonding π orbitals. (B) All of its electrons are paired in bonding orbitals. (C) Its bond order is zero. (D) It has six bonding regions around a central atom.

2. **True or false.** A molecular-orbital bond order of zero predicts a stably bound molecule in that configuration.

3. **Drag the words.** Complete the molecular descriptions. A bond order is half the difference between _____. Six bonding regions with no lone pairs give _____ geometry. Vibrational spectra are mainly _____, while rotational spectra are mainly _____. (Terms: bonding and antibonding electron counts; octahedral; infrared; microwave; tetrahedral; ultraviolet only.)

4. **Fill in the blanks.** Complete the bond-order and geometry results. If a molecule has 8 bonding and 4 antibonding electrons, its bond order is _____; six electron regions with no lone pairs form an _____ geometry.

5. **Mark the words.** Mark the two intermolecular-force statements. “London dispersion acts between all atoms and molecules, while especially strong hydrogen bonding requires hydrogen bonded to nitrogen, oxygen, or fluorine.”

12.6 Conceptual Questions

1. Contrast valence-bond and molecular-orbital descriptions of where bonding electrons reside.

2. Why does filling an antibonding orbital reduce bond order?

3. Oxygen is paramagnetic. Why is this evidence for the molecular-orbital description?

4. Explain why electron-domain geometry, molecular geometry, and hybridization are related but not identical descriptions.

5. Why does hydrogen bonding strongly raise water’s boiling point compared with hydrogen sulfide?

6. Why does a diatomic molecule’s infrared spectrum contain many rotational lines around each vibrational transition?

12.7 Problems

Exercise 12.1

(Straightforward)

Determine the hybridization of the central atom and predict the molecular geometry for (a) NH_3 (three bonding pairs, one lone pair on N), (b) CO_2 (two double bonds, no lone pairs on C), (c) SF_6.

Exercise 12.2

(Straightforward)

Carbon monoxide, CO, is isoelectronic with N_2: it has the same number of valence electrons (4 from C, 6 from O, 10 total) and, being built from two light second-row atoms, fills its molecular orbitals in the same order used for N_2 above: $\sigma_{2s}, \sigma^*_{2s}, \pi_{2p}$ ($\times 2$), σ_{2p}. Construct the molecular orbital diagram for CO, determine its bond order, and compare it to the triple bond expected from the Lewis structure : C$\equiv$O :.

Exercise 12.3

(Straightforward)

Using the same style of reasoning applied to H_2 and He_2 in the text, determine the bond order predicted by MO theory for the hypothetical ion He_2^+ (three electrons: two in σ_{1s}, one in σ^*_{1s}), and state whether this ion is predicted to be (marginally) stable.

Exercise 12.4

(Straightforward)

The HCl molecule has an effective vibrational frequency $f = \omega/2\pi = 8.66 \times 10^{13}$ Hz. Using $m_{\mathrm{H}} = 1.008$ u and $m_{\mathrm{Cl}} = 35.45$ u (1 u $= 1.66 \times 10^{-27}$ kg), compute (a) the reduced mass μ, and (b) the zero-point vibrational energy $E_0 = \frac{1}{2}\hbar\omega$ in eV.

Exercise 12.5

(Moderate)

The CO molecule has bond length $r_0 = 0.113$ nm and reduced mass $\mu = 6.86$ u. Compute (a) its moment of inertia $I = \mu r_0^2$, and (b) the energy (in units of 10^{-4} eV) of the $J = 0 \to J = 1$ rotational transition.

Exercise 12.6

(Moderate)

Explain, using the concept of bond order, why O_2^- (superoxide, one more electron than O_2) has a weaker, longer bond than neutral O_2, while O_2^+ (dioxygenyl, one fewer electron) has a stronger, shorter bond — referring to which type of orbital (bonding or antibonding) the added or removed electron occupies.

Exercise 12.7

(Moderate)

Determine the hybridization of the central atom and predict the molecular geometry of SF_4 (four bonding pairs and one lone pair on S). Sketch, in words, where the lone pair sits relative to the five electron domains of the underlying trigonal-bipyramidal arrangement, and state the resulting molecular geometry. Compare your reasoning to the XeF_4 worked example above: why does one lone pair produce a very different-looking molecule than two lone pairs do?

Exercise 12.8
(Moderate)

Using the same reasoning applied to N_2/N_2^+ in the text and to $O_2/O_2^\pm$ in Exercise 12.6, construct the molecular orbital diagram for neutral F_2 (14 valence electrons, heavier-diatomic ordering: $\sigma_{2s}, \sigma^*_{2s}, \sigma_{2p}, \pi_{2p}\,(\times 2), \pi^*_{2p}\,(\times 2), \sigma^*_{2p}$), and determine its bond order. Then remove one electron to form F_2^+, identify which orbital it comes from, and state whether the ion's bond is predicted to be stronger or weaker (and shorter or longer) than neutral F_2's, and whether the ion is paramagnetic or diamagnetic.

Exercise 12.9
(Challenging)

Using the reduced mass μ found in Exercise 12.4(a) for HCl and a bond length $r_0 = 127.5$ pm, find (a) the moment of inertia I, (b) the rotational constant B in eV, (c) the rovibrational line spacing $2B$ in cm^{-1}, and (d) the size of the gap (in cm^{-1}) at the band origin left by the missing Q branch.

Exercise 12.10
(Challenging)

The boiling points of the halogens rise steadily down the group: F_2, -188°C; Cl_2, -34°C; Br_2, 59°C; I_2, 184°C. All four are nonpolar diatomic molecules with zero permanent dipole moment. Explain this trend using London dispersion forces, and state what physical property of the molecule is chiefly responsible for it.

Part V

Nuclear and Particle Physics

Chapter 13

Nuclear Physics

Learning Objectives

By the end of this chapter, you should be able to:

- Describe the composition of the nucleus and use standard notation for nuclides and isotopes.
- Compute nuclear radius, density, binding energy, and binding energy per nucleon, and interpret the binding-energy-per-nucleon curve.
- Explain the qualitative form of the strong nuclear force and the semi-empirical (liquid-drop) mass formula.
- Apply the exponential radioactive decay law, including half-life, mean lifetime, and activity.
- Distinguish alpha, beta, and gamma decay, explain tunneling's role in alpha decay, and apply conservation laws to each.
- Explain nuclear fission and fusion in terms of the binding-energy curve and compute energy released in representative reactions.
- Write the semi-empirical (Weizsäcker) mass formula term by term, and use it to explain the extra stability of even-even nuclei and of the "magic-number" nuclides predicted by the nuclear shell model.
- Trace a multi-step radioactive decay chain, apply the decay law to isotopic (radiocarbon) dating, and sketch the quantitative dependence of alpha-decay tunneling probability on barrier height and width.

Introduction

Chapters 10–12 treated the nucleus as a structureless positive point charge, a fixed source of the Coulomb potential binding atomic electrons. This chapter looks inside the nucleus itself: a bound system of protons and neutrons, held together not by the electromagnetic force (which, among the mutually repelling protons, would tend to blow the nucleus apart) but by a new fundamental interaction, the strong nuclear force, with a strength and range entirely different from anything encountered so far in this book. The same quantum-mechanical ideas developed for atoms — quantized energy levels, tunneling, exponential decay driven by fixed transition probabilities — reappear here on a length scale roughly 10^5 times smaller and an energy scale roughly 10^6 times larger, and account for radioactivity, nuclear stability, and the energy-release mechanisms (fission and fusion) that make the nucleus, unlike the atom, a practical source of usable energy.

13.1 Nuclear Structure and Binding Energy

Nuclear Composition and Notation

A nucleus consists of Z protons and N neutrons, collectively called **nucleons**, with mass number $A = Z + N$. A given nuclear species (**nuclide**) is denoted ${}^{A}_{Z}X$, where X is the chemical symbol determined by Z (since Z alone fixes the number of atomic electrons in the neutral atom, and hence its chemistry). Nuclides sharing the same Z but different N (and hence different A) are **isotopes** of the same element — chemically near-identical but differing in mass and, often, nuclear stability. Protons and neutrons have nearly equal mass ($m_p c^2 = 938.3$ MeV, $m_n c^2 = 939.6$ MeV), and because a proton's charge is exactly opposite an electron's, the notation ${}^{A}_{Z}X$ carries the atom's full identity without needing N written explicitly ($N = A - Z$).

The division of labor between Z and N — one fixes the chemistry, the other only the mass and the stability — is worth handling directly before the rest of the chapter leans on it. In Figure 13.1, adding a neutron to an atom leaves the element name untouched and moves along a row of isotopes, while adding a proton changes the element outright. The simulation also assembles the tabulated atomic mass of an element from its isotopes weighted by natural abundance, which is why the periodic table lists 35.45 for chlorine although no chlorine nucleus has that mass.

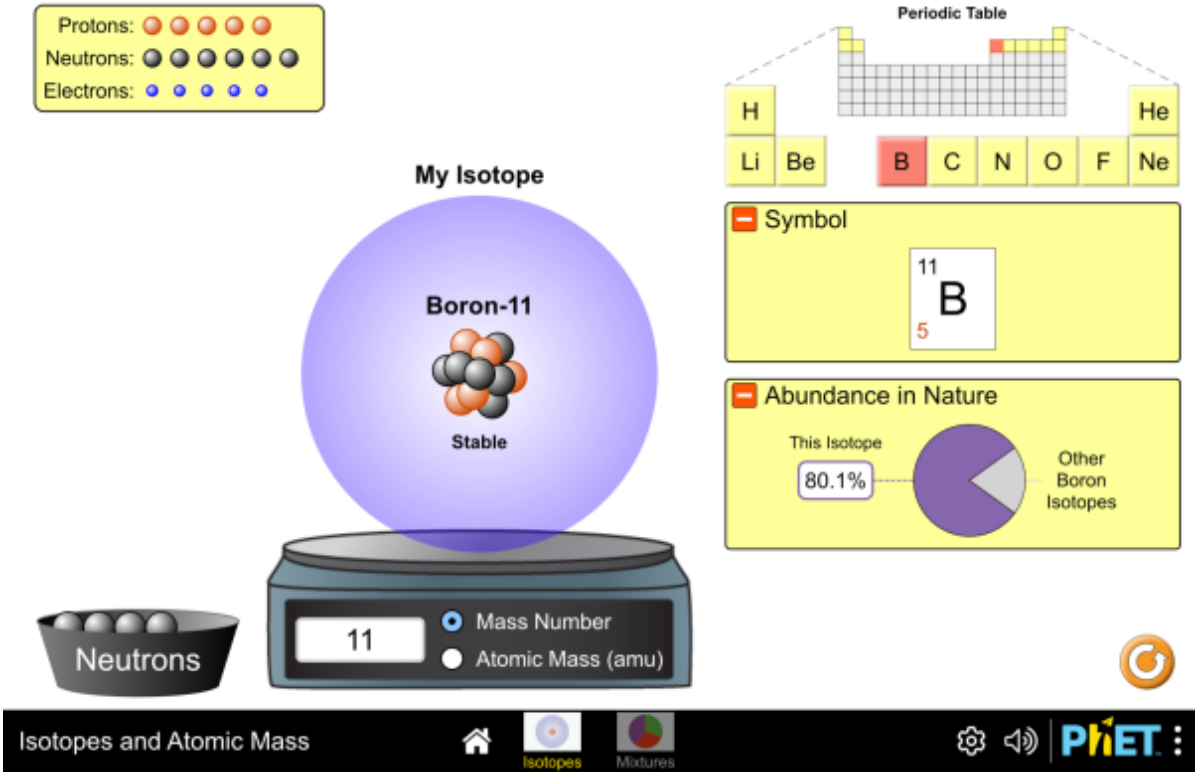

Figure 13.1: Isotopes built nucleon by nucleon, with natural abundances and the resulting average atomic mass. Z names the element; N decides how long the nucleus lasts. Interactive simulation: Isotopes and Atomic Mass

Nuclear Size, Density, and the Strong Force

Scattering experiments (extending Rutherford's original alpha-scattering method, now generally using higher-energy electron or nucleon probes to resolve the nuclear interior itself) show that nuclear radius grows with mass number as

$$R = R_0 A^{1/3}, \qquad R_0 \approx 1.2 \text{ fm } (1 \text{ fm} = 10^{-15} \text{ m}), \tag{13.1}$$

so nuclear *volume* is proportional to A — each nucleon occupies, on average, the same volume regardless of the size of the nucleus it belongs to, exactly as one would expect for an (nearly) incompressible fluid of tightly packed, closely spaced constituents. This is the empirical basis of the **liquid-drop model** of the nucleus, in which the nucleus is treated, for many purposes, as a droplet of incompressible nuclear fluid.

The corresponding mass density is therefore nearly independent of A:

$$\rho \approx \frac{A m_n}{(4\pi/3)R^3} = \frac{3m_n}{4\pi R_0^3} \approx 2.3 \times 10^{17} \text{ kg/m}^3 \approx 0.14 \text{ nucleons/fm}^3. \tag{13.2}$$

This enormous but nearly constant density is one of the clearest signs that nuclei behave like an incompressible nuclear fluid rather than like dilute collections of independently orbiting particles.

Nuclear stability is not explained by electromagnetism — the Coulomb force between two protons at nuclear separations, ~ 1 fm, is enormously repulsive and would blow the nucleus apart if the electromagnetic force between nucleons were the whole story. Nuclei are held together by the **strong nuclear force**, an attractive interaction between nucleons at typical nuclear separations (proton-proton, proton-neutron, and neutron-neutron alike, largely independent of charge) that is far stronger than the Coulomb repulsion at nuclear distances. It has an extremely short range (roughly 1–2 fm), falling off essentially to zero beyond a few fermis; at still shorter separations, the nuclear interaction also has a repulsive core that prevents the nucleus from collapsing. This short range explains why nuclear binding, unlike Coulomb binding, saturates: a given nucleon interacts strongly only with its immediate neighbors, not with every other nucleon in the nucleus (unlike the long-range Coulomb repulsion, which every proton feels from every other proton, growing roughly as Z^2) — a key qualitative fact used below to explain both the shape of the binding-energy curve and, ultimately, nuclear fission and fusion.

> For scale, 1 fm $= 10^{-15}$ m is about 10^5 times smaller than a typical atomic radius (~ 1 Å $= 10^{-10}$ m) — the same factor of 10^5 flagged in the chapter introduction as the length scale separating nuclear from atomic physics.

The measurement that started all of this is reconstructed in Figure 13.2. Alpha particles are fired at a nucleus of adjustable Z, and almost all of them pass through with a barely perceptible deflection — but a few come back. Rutherford's inference from that rare, large-angle scattering was that the positive charge is concentrated in a volume tiny compared with the atom, and the simulation makes the inference reproducible: shrink the nuclear charge and the sharp back-scattering disappears, while the plum-pudding screen shows what a diffuse charge distribution would have produced instead — no large deflections at all, at any impact parameter.

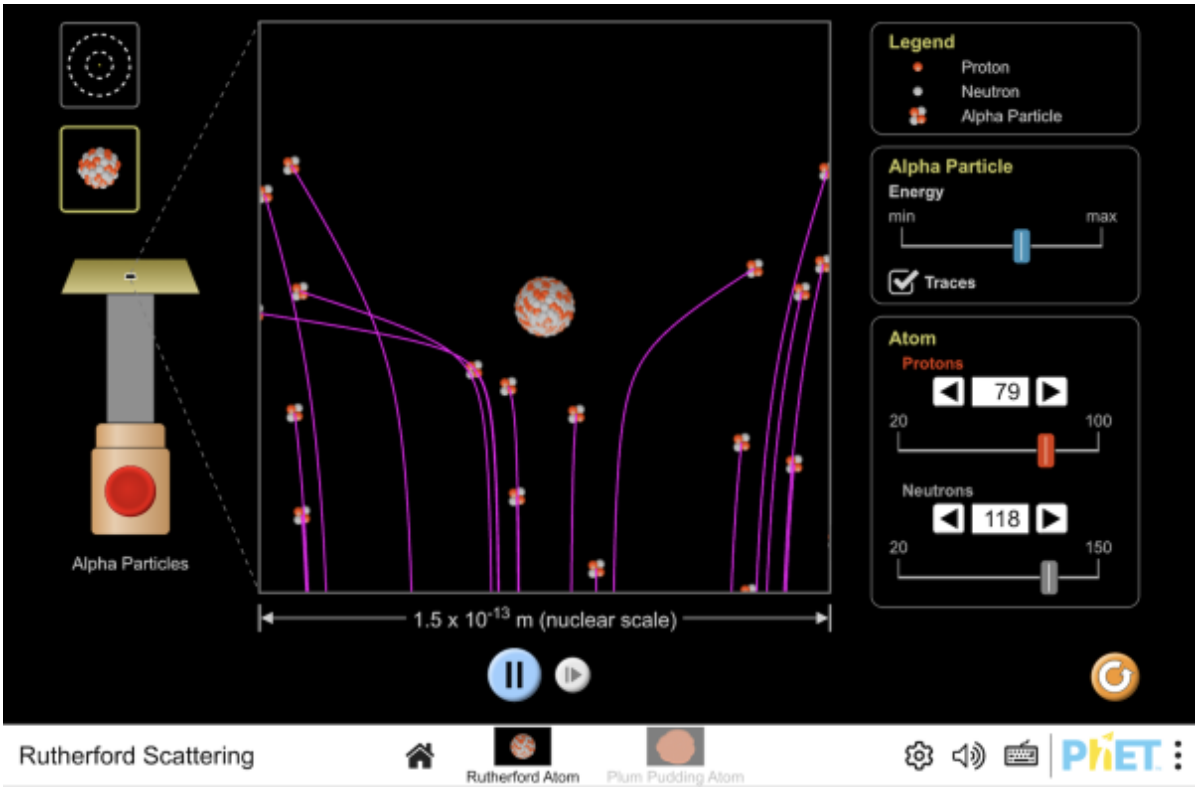

Figure 13.2: Alpha particles scattering from a nucleus of adjustable charge, with a diffuse-charge atom available for comparison. The trajectories are pure Coulomb hyperbolas; the strong force never enters, which is exactly why the experiment measures charge and size rather than nuclear structure. Interactive simulation: Rutherford Scattering

Binding Energy

The mass of a bound nucleus is always *less* than the sum of the masses of its separated constituent protons and neutrons — a direct manifestation of mass–energy equivalence (Chapter 3): energy must be supplied to pull the nucleus apart into free nucleons, so the bound system, having lower total energy, has correspondingly lower total mass. In terms of **nuclear** masses, the **binding energy** is defined as

$$E_B = \left[Zm_p + Nm_n - M_{\text{nuc}}({}^{A}_{Z}X)\right] c^2, \tag{13.3}$$

where $M_{\text{nuc}}({}^{A}_{Z}X)$ is the nuclear mass. Since mass tables usually give **atomic** masses, the equivalent form that can be evaluated directly from those tables is

$$E_B = \left[Zm({}^{1}\text{H}) + Nm_n - M_{\text{atom}}({}^{A}_{Z}X)\right] c^2, \tag{13.4}$$

neglecting the tiny differences in atomic electron-binding energies. The hydrogen-atom mass replaces the bare proton mass so that the Z electron masses cancel. It is generally more informative to consider the **binding energy per nucleon**, E_B/A, since this measures how tightly, on average, an individual nucleon is bound, independent of the nucleus's overall size.

> The **atomic mass unit** is defined as exactly 1/12 the mass of a neutral ${}^{12}_{6}\text{C}$ atom, $1\ \text{u} = 1.660539 \times 10^{-27}$ kg, converting to energy as $1\ \text{u}\,c^2 = 931.494$ MeV — the conversion factor used throughout this chapter's worked examples.

Plotting E_B/A against A for all known nuclides gives one of the most important curves in nuclear physics: E_B/A rises sharply from very light nuclei, peaks at around $E_B/A \approx 8.7$ MeV near $A \approx 56$ (iron and its neighbors), and then decreases slowly for heavier nuclei. This shape is easiest to absorb by looking at it directly, plotted from real nuclear data, in Figure 13.3.

> The single most tightly bound nuclide per nucleon is actually ${}^{62}_{28}\text{Ni}$, not ${}^{56}_{26}\text{Fe}$, though the two are close enough (8.7946 vs. 8.7903 MeV/nucleon) that "the iron peak" remains the standard shorthand for this region of the curve.

Two competing effects, both traceable to the short range of the strong force versus the long range of the Coulomb force, explain this shape: for light nuclei, a growing fraction of nucleons sit at or near the nuclear surface, with fewer strong-force neighbors than an interior nucleon has (a **surface term**, reducing E_B/A for small A, since surface-to-volume ratio falls as A grows), while for heavy nuclei, the number of proton pairs, and hence the total Coulomb repulsion energy, grows roughly as Z^2 — much faster than the (short-range, saturating) strong-force binding, which grows only as A — progressively weakening binding per nucleon as A increases (a **Coulomb term**). Together with a bulk (volume) term that would alone give constant E_B/A, and additional smaller terms accounting for the extra stability of nuclei with $N = Z$ (a symmetry term) and of nuclei with even numbers of protons and neutrons (a pairing term), these considerations make up the **semi-empirical mass formula** (also called the Weizsäcker or liquid-drop mass formula), which reproduces the observed binding energy of essentially every known nuclide to good accuracy using only five fitted terms with a clear physical origin apiece.

The location of the peak at $A \approx 56$ has an immediate and far-reaching consequence, developed further below: **energy can be released either by combining light nuclei into**

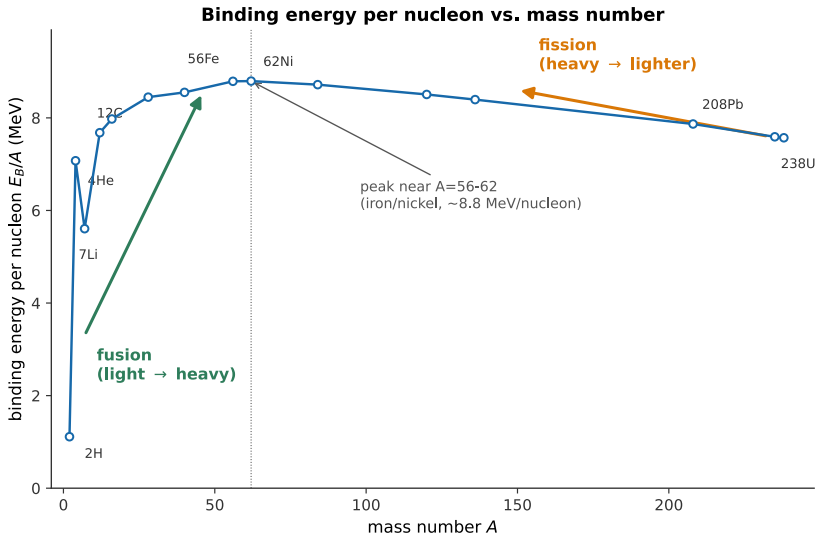

Figure 13.3: Binding energy per nucleon vs. mass number, computed from standard atomic mass data for a representative set of nuclides from deuterium to uranium-238. The peak, at $A \approx 56$–62 (iron and nickel), is the single most consequential feature of the curve: fusion climbs the steep left flank toward it, and fission descends the shallow right flank toward it, both releasing energy in the process, as developed below. Original figure by the author, computed with matplotlib from standard nuclear mass data.

heavier ones (fusion) or by splitting heavy nuclei into lighter ones (fission), in each case moving the participating nucleons toward the peak of the curve, where they are more tightly bound (lower mass) than before — with the energy released equal to the resulting decrease in total rest mass, via $E = \Delta m\, c^2$.

The Semi-Empirical Mass Formula

Written out in full, with the volume, surface, Coulomb, symmetry, and pairing terms named above each given an explicit form, the semi-empirical (or Weizsäcker) mass formula predicts the binding energy of a nuclide with Z protons and mass number A ($N = A - Z$ neutrons) as

$$E_B(Z,A) = \underbrace{a_V A}_{\text{volume}} - \underbrace{a_S A^{2/3}}_{\text{surface}} - \underbrace{a_C \frac{Z(Z-1)}{A^{1/3}}}_{\text{Coulomb}} - \underbrace{a_A \frac{(A-2Z)^2}{A}}_{\text{symmetry}} + \underbrace{\delta(A,Z)}_{\text{pairing}}, \tag{13.5}$$

with, in one standard fit to measured masses (due to Krane),

$$a_V = 15.5 \text{ MeV}, \quad a_S = 16.8 \text{ MeV}, \quad a_C = 0.72 \text{ MeV}, \quad a_A = 23 \text{ MeV}, \tag{13.6}$$

$$\delta(A,Z) = \begin{cases} +a_P A^{-3/4}, & Z \text{ and } N \text{ both even (even–even)} \\ 0, & A \text{ odd} \\ -a_P A^{-3/4}, & Z \text{ and } N \text{ both odd (odd–odd)} \end{cases}, \qquad a_P \approx 34 \text{ MeV}. \tag{13.7}$$

(Different published fits to the mass table give somewhat different coefficients — a reminder that this is a phenomenological model fit to data, not derived from first principles — but all agree on the same five terms and roughly the same magnitudes.) Each term has already been

motivated qualitatively above: the **volume term** $a_V A$ reflects the short-range, saturating strong force acting equally on every nucleon, as if each contributed a fixed amount of binding independent of A; the **surface term** $-a_S A^{2/3}$ subtracts binding from the nucleons at the nuclear surface (proportional to surface area, $\propto R^2 \propto A^{2/3}$), which have fewer strong-force neighbors than an interior nucleon; the **Coulomb term** $-a_C Z(Z-1)/A^{1/3}$ is the electrostatic self-energy of Z mutually repelling protons packed into a sphere of radius $R \propto A^{1/3}$, reducing binding as the number of proton pairs, $\sim Z^2$, grows.

The **symmetry term** $-a_A(A-2Z)^2/A$ is new here: it penalizes any departure from $N = Z$, and follows from the Pauli exclusion principle rather than from either the strong or Coulomb force directly — protons and neutrons fill separate sets of momentum states (each a distinct kind of fermion, Chapter 11), so for fixed A, forcing N and Z apart pushes some nucleons into higher-energy states than a balanced $N = Z$ filling would require, exactly as forcing electrons into higher shells costs energy in a many-electron atom. Maximizing E_B with respect to Z at fixed A (equivalently, minimizing the nuclear mass) gives, for large A,

$$Z_{\min}(A) \approx \frac{2a_A A}{4a_A + a_C A^{2/3}}, \tag{13.8}$$

the proton number that maximizes binding energy for that A. For $A = 238$, this predicts $Z_{\min} \approx 91.5$ — strikingly close to uranium's actual $Z = 92$ — and, more generally, shows why the ratio $Z_{\min}/A$ decreases steadily below $1/2$ as A grows: the Coulomb term's Z^2 growth increasingly outweighs the symmetry term's preference for $N = Z$, pushing the most stable nucleus for each A toward greater neutron excess, exactly the neutron-rich bending of the stable band away from $N = Z$ seen in Figure 13.8 below.

Deriving the Line of Stability: Maximizing E_B at Fixed A

Only the Coulomb and symmetry terms depend on how a fixed A is divided between Z and N; the volume, surface, and (for this smooth, continuous-Z treatment) pairing terms are dropped. Writing E_B as a function of Z alone at fixed A and differentiating,

$$\frac{\partial E_B}{\partial Z} = -a_C \frac{2Z-1}{A^{1/3}} + \frac{4a_A(A-2Z)}{A}, \tag{13.9}$$

using $\partial[Z(Z-1)]/\partial Z = 2Z-1$ and $\partial[(A-2Z)^2]/\partial Z = -4(A-2Z)$. Setting this to zero and approximating $2Z-1 \approx 2Z$ (an excellent approximation once Z is more than a few units) gives

$$\frac{2a_C Z}{A^{1/3}} = \frac{4a_A(A-2Z)}{A}. \tag{13.10}$$

Multiplying both sides by A and collecting every term proportional to Z on the left,

$$a_C Z A^{2/3} + 4a_A Z = 2a_A A \quad \Longrightarrow \quad Z\left(4a_A + a_C A^{2/3}\right) = 2a_A A, \tag{13.11}$$

which rearranges directly into the formula quoted in the text,

$$Z_{\min}(A) = \frac{2a_A A}{4a_A + a_C A^{2/3}}. \tag{13.12}$$

Because this comes from setting a first derivative to zero, it locates a maximum of E_B (equivalently, a minimum of the nuclear mass) at fixed A — exactly the most stable nuclide for that mass number, which is what "the line of stability" means.

The **pairing term** $\delta(A, Z)$ captures a purely quantum effect with no analog in the volume, surface, or Coulomb terms: nucleons of the same type couple preferentially in spin-paired twos (much as the two electrons of a filled atomic orbital pair their spins, Chapter 11), each pair contributing extra binding beyond what the smooth terms above predict. An **even–even** nuclide (even Z, even N) has every nucleon paired and gains the full pairing bonus; an **odd–odd** nuclide has one unpaired proton and one unpaired neutron and loses it; an odd-A nuclide (one odd, one even) is intermediate and conventionally assigned $\delta = 0$. The consequence is stark and directly observable in the chart of the nuclides: of the roughly 260 stable nuclides, about 150 are even–even, roughly 100 are odd-A, and only four are odd–odd ($^{2}_{1}$H, $^{6}_{3}$Li, $^{10}_{5}$B, and $^{14}_{7}$N, all among the very lightest nuclei, where the Coulomb penalty for the proton-rich $N = Z$ arrangement is still modest) — essentially every heavier odd–odd combination is unstable, decaying by beta emission toward an even–even or odd-A neighbor of lower mass.

Worked Example: Binding Energy from the Semi-Empirical Mass Formula

Apply the semi-empirical mass formula to $^{56}_{26}$Fe ($A = 56$, $Z = 26$, $N = 30$, even–even) and compare to the measured value. Term by term,

$$a_V A = 15.5(56) = 868.0 \text{ MeV}, \qquad a_S A^{2/3} = 16.8(56)^{2/3} = 245.9 \text{ MeV}, \tag{13.13}$$

$$a_C \frac{Z(Z-1)}{A^{1/3}} = 0.72\,\frac{26(25)}{56^{1/3}} = 122.3 \text{ MeV}, \qquad a_A \frac{(A-2Z)^2}{A} = 23\,\frac{4^2}{56} = 6.6 \text{ MeV}, \tag{13.14}$$

and, since ^{56}Fe is even–even, the pairing bonus $\delta = +a_P A^{-3/4}$,

$$\delta = 34(56)^{-3/4} \approx 1.7 \text{ MeV}. \tag{13.15}$$

Summing with the signs shown in the formula above,

$$E_B \approx 868.0 - 245.9 - 122.3 - 6.6 + 1.7 = 494.9 \text{ MeV}, \qquad E_B/A \approx 8.84 \text{ MeV/nucleon}. \tag{13.16}$$

The measured value, computed from $^{56}_{26}$Fe's tabulated atomic mass $M = 55.934936$ u using $E_B = \left[Z\,m(^1\text{H}) + Nm_n - M\right]c^2$ (with $m(^1\text{H}) = 1.007825$ u, the *atomic* hydrogen mass, so that the electron masses on both sides of the equation cancel exactly rather than merely approximately), is $E_B = 492.3$ MeV, or 8.79 MeV/nucleon. The five-term fit reproduces the measured binding energy to about half a percent — remarkable accuracy from a formula built on the bulk liquid-drop picture of the nucleus plus one quantum correction, and a large part of why the semi-empirical mass formula remains a standard first tool for estimating the mass of a nuclide not yet directly measured.

The Nuclear Shell Model and Magic Numbers

The semi-empirical mass formula treats the nucleus as a featureless liquid drop and, term by term, does remarkably well — but it is not the whole story. Certain nuclides, with Z or N equal to one of the **magic numbers** $2, 8, 20, 28, 50, 82, 126$, are measurably more tightly bound, more abundant, and more resistant to further reaction (larger energy gap to the first excited state, smaller neutron-capture cross section) than the smooth semi-empirical formula predicts — $^{4}_{2}$He, $^{16}_{8}$O, $^{40}_{20}$Ca, and $^{208}_{82}$Pb (with $N = 126$) all sit at pronounced local peaks of the true

binding-energy surface, and "doubly magic" nuclides such as $^{208}_{82}$Pb (magic in both Z and N) are exceptionally stable for their mass. $Z = 50$ (tin) is a magic proton number too: tin has ten stable isotopes, more than any other element, a record set jointly by its magic-number proton-shell closure and (since $Z = 50$ is even) the full pairing bonus of an even-Z nuclide — in sharp contrast to its odd-Z neighbor antimony ($Z = 51$), which has only two. This is the nuclear analog of the noble-gas closed-shell stability seen in atomic electron structure (Chapter 11): just as an atom with a filled electron shell ($Z = 2, 10, 18, \ldots$) is unusually inert, a nucleus with a filled proton or neutron shell is unusually well bound, for the same underlying reason — a large energy gap separates the top of a filled shell from the next available single-particle state, so removing or adding one more nucleon costs (or gains) far more energy than for a nucleus with a partially filled shell.

The **nuclear shell model** treats each nucleon, to a first approximation, as moving independently in an average central potential produced by all the other nucleons combined — conceptually similar to the mean-field treatment used for atomic electrons (Chapter 11), though for a very different potential shape. As previewed in Chapter 9, the isotropic three-dimensional harmonic oscillator is a convenient starting point for this average potential (its energy levels and degeneracies are exactly solvable), but the oscillator's own degeneracies alone reproduce only the lightest magic numbers $(2, 8, 20)$; reproducing the full sequence up to 126 requires an additional **spin-orbit coupling** term (splitting each oscillator level by the nucleon's intrinsic spin coupled to its orbital motion, an effect far stronger, relative to the level spacing, than the analogous fine-structure splitting in atoms), first added by Maria Goeppert Mayer and, independently, J. Hans D. Jensen in 1949 — work for which they shared the 1963 Nobel Prize in Physics. With the spin-orbit term included, the shell model correctly reorders the single-particle levels so that the cumulative count of states filled at each major shell closure lands exactly on the observed magic numbers, explaining not only the extra binding at magic Z or N but also finer details such as the nuclear spins and magnetic moments of odd-A nuclides (set, in the simplest version of the model, entirely by the single unpaired nucleon outside an otherwise paired, spherically symmetric core).

13.2 Radioactive Decay

An unstable nuclide decays into a different nuclide (or a lower energy state of the same nuclide) at a rate governed, as with any quantum system undergoing a transition, by a fixed, per-nucleus **decay constant** λ (units of inverse time), independent of the nucleus's history or environment (with rare, small exceptions for certain electron-capture processes sensitive to chemical environment). If $N(t)$ is the number of undecayed nuclei present at time t, the rate of decay is proportional to the number remaining, $dN/dt = -\lambda N$, giving the **exponential decay law**:

$$N(t) = N_0\, e^{-\lambda t}. \tag{13.17}$$

The **half-life** $T_{1/2}$, the time for half of an initial sample to decay, and the **mean lifetime** τ, the average lifetime of an individual nucleus, are related to λ by

$$T_{1/2} = \frac{\ln 2}{\lambda}, \qquad \tau = \frac{1}{\lambda} = \frac{T_{1/2}}{\ln 2}. \tag{13.18}$$

A Half-Life Is Not a Countdown Timer

It is tempting to picture a radioactive nucleus as "aging" toward its half-life, becoming more likely to decay the longer it has already survived — as if $T_{1/2}$ were a fuse burning down. It is not: the decay constant λ is fixed and history-independent, so a nucleus that has already survived ten half-lives is exactly as likely to decay in the next second as one created a moment ago. A related error is treating $T_{1/2}$ and the mean lifetime τ as interchangeable — they are not. $\tau = T_{1/2}/\ln 2 \approx 1.44\,T_{1/2}$ is *longer* than the half-life, because the relatively rare nuclei that survive far longer than typical pull the average up, even though the *median* survival time is exactly $T_{1/2}$.

The **activity**, $\mathcal{A} \equiv -dN/dt = \lambda N(t) = \mathcal{A}_0 e^{-\lambda t}$, is the physically measured decay rate (in decays per second, or the traditional unit the curie), and decays with the same exponential form and the same $T_{1/2}$ as $N(t)$ itself, since $\mathcal{A}$ is simply proportional to N at every instant.

1 curie (Ci) $\equiv 3.7 \times 10^{10}$ decays/s, originally chosen to approximate the activity of 1 g of ^{226}Ra (verified explicitly in the worked example below). The SI unit is the **becquerel** (Bq), 1 Bq $\equiv$ 1 decay/s.

The exponential law describes a population, not a nucleus, and the difference between those two statements is where intuition usually fails. Figure 13.4 shows both at once: individual nuclei decaying at unpredictable moments with no memory of how long they have already waited, and the count of survivors nevertheless tracing a clean exponential once the sample is large enough. The same simulation puts the law to work in reverse — measuring the residual ^{14}C or ^{238}U fraction in a sample of unknown age and reading off the elapsed time — which is the argument by which the age of a bone, a lava flow, or the Earth itself is established.

Historical Context: Becquerel, the Curies, and the Discovery of Radioactivity

Radioactivity was discovered by accident. In early 1896, prompted by Wilhelm Röntgen's announcement of X-rays only weeks before, Henri Becquerel was investigating whether phosphorescent materials (which glow after exposure to light) might also emit penetrating, photographic-plate-fogging radiation, using uranium salts as his test material. An overcast Paris sky in late February 1896 forced him to store an unexposed sample — uranium salt sitting atop a wrapped photographic plate, with no sunlight available to trigger any phosphorescence — in a dark drawer for several days. When he developed the plate anyway on March 1, expecting at most a faint trace, he found a strong image instead: the uranium was emitting penetrating radiation entirely on its own, with no light exposure needed to trigger it. Becquerel had discovered **radioactivity**, a spontaneous nuclear process, without initially recognizing what he had found (he first suspected an unusually persistent, invisible form of phosphorescence). Becquerel himself is shown in a portrait from around the time of the discovery in Figure 13.6.

Marie Skłodowska Curie, joined by her husband Pierre, took up the puzzle as a doctoral research topic and pushed it decisively further. Testing pitchblende ore (the raw uranium mineral), she found it *more* radioactive than its uranium content alone could account for — direct evidence of one or more additional, still-unidentified radioactive elements hidden within it in trace quantities. Over four years of laborious chemical separation, processing tons of ore

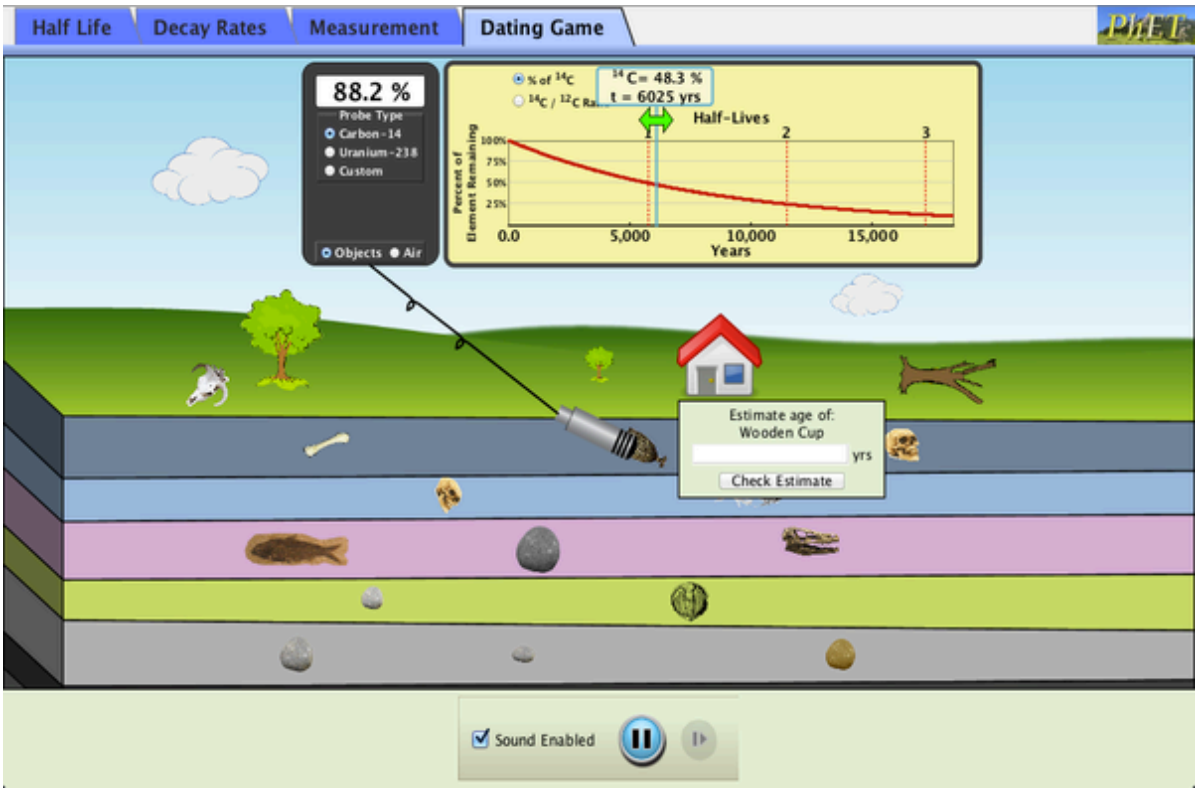

Figure 13.4: Radioactive decay watched nucleus by nucleus and in bulk, then applied to dating. The decay constant is a probability per unit time and nothing else: nuclei do not age. Interactive simulation: Radioactive Dating Game



Figure 13.5: A plain-language reminder of the distinction this section relies on throughout: it is the decaying nuclei, not "radiation" as some free-floating substance, that the decay law above actually describes. Watch on YouTube: Radiation vs Radioactive Atoms

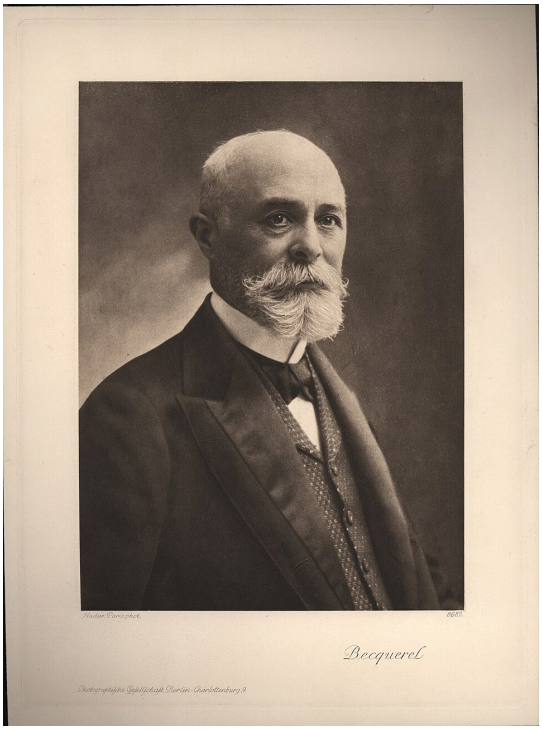

Figure 13.6: Henri Becquerel (1852–1908). Photograph by Paul Nadar, before 1908; Smithsonian Institution Libraries (Dibner Library collection); public domain in the United States via Wikimedia Commons.

by hand in a converted shed, the Curies isolated two new elements, **polonium** (1898, named for Marie's native Poland) and **radium** (1898), and coined the term **radioactivity** itself for the phenomenon. Becquerel and the Curies shared the 1903 Nobel Prize in Physics for this work; Marie Curie went on to win a second Nobel Prize, in Chemistry, in 1911, for the isolation and characterization of radium and polonium as new chemical elements — achievements made decades before the neutron or the proton-neutron picture of the nucleus used throughout this chapter was available to explain, even in outline, what radioactivity actually *was*.

A Decay Chain: Uranium-238 to Lead-206

A single alpha or beta decay usually does not, by itself, produce a stable nuclide: the daughter is often itself radioactive, decaying again, and again, until a stable nuclide is finally reached. Such a sequence is a **decay chain** (or decay series). The longest and most geologically important chain starts at ${}^{238}_{92}\mathrm{U}$ and ends, fourteen decays later, at stable ${}^{206}_{82}\mathrm{Pb}$, through a sequence of alpha and beta decays; its first several steps are

$$ {}^{238}_{92}\mathrm{U} \xrightarrow{\alpha} {}^{234}_{90}\mathrm{Th} \xrightarrow{\beta^-} {}^{234}_{91}\mathrm{Pa} \xrightarrow{\beta^-} {}^{234}_{92}\mathrm{U} \xrightarrow{\alpha} {}^{230}_{90}\mathrm{Th} \xrightarrow{\alpha} {}^{226}_{88}\mathrm{Ra} \xrightarrow{\alpha} {}^{222}_{86}\mathrm{Rn} \xrightarrow{\alpha} \cdots \longrightarrow {}^{206}_{82}\mathrm{Pb}. \tag{13.19} $$

In all, the full chain from ${}^{238}\mathrm{U}$ to ${}^{206}\mathrm{Pb}$ involves eight alpha decays and six beta-minus decays — a bookkeeping check confirms it must: eight alpha decays remove $8 \times 4 = 32$ from the mass number ($238 - 32 = 206$, matching ${}^{206}\mathrm{Pb}$ exactly) and $8 \times 2 = 16$ from the atomic number, while six beta-minus decays each raise Z by one without changing A, for a net $\Delta Z = -16 + 6 = -10$ ($92 - 10 = 82$, matching lead exactly). Each step has its own half-life,

ranging from the parent's 4.5×10^9 years (comparable to the age of the solar system, which is why primordial ^{238}U is still present in Earth's crust at all) down to a fraction of a second for the shortest-lived member of the chain — a span of many orders of magnitude in half-life among nuclides connected by nothing more exotic than successive alpha and beta emissions.

Because every step in a chain like this decays at its own fixed rate while simultaneously being *replenished* by the decay of its parent, a chain that has run undisturbed for a time long compared to every half-life except the first settles into **secular equilibrium**: each intermediate's activity (λN for that nuclide) becomes equal to the activity of the slowly decaying parent feeding it, even though the intermediate's own N can be wildly different from the parent's. This is why an old, undisturbed sample of uranium ore contains measurable, steady quantities of highly radioactive but short-lived daughters such as radium and radon: they are being produced by ^{238}U decay just as fast as they themselves decay away.

Worked Example: Activity of a Radium-226 Sample

$^{226}_{88}$Ra (half-life $T_{1/2} = 1600$ years), a step in the uranium-238 chain above, is the isotope Marie Curie isolated and the one for which the **curie** unit of activity was originally defined: 1 Ci $\equiv 3.7 \times 10^{10}$ decays/s, chosen to approximate the activity of 1 g of ^{226}Ra. Verify this using the decay law. The number of nuclei in 1.00 g of ^{226}Ra (molar mass ≈ 226 g/mol) is

$$N = \frac{(1.00\ \text{g})(6.022 \times 10^{23}\ \text{mol}^{-1})}{226\ \text{g/mol}} = 2.665 \times 10^{21}, \tag{13.20}$$

and the decay constant is

$$\lambda = \frac{\ln 2}{T_{1/2}} = \frac{0.6931}{(1600\ \text{yr})(3.156 \times 10^{7}\ \text{s/yr})} = 1.372 \times 10^{-11}\ \text{s}^{-1}. \tag{13.21}$$

The activity is then

$$\mathcal{A} = \lambda N = (1.372 \times 10^{-11}\ \text{s}^{-1})(2.665 \times 10^{21}) = 3.66 \times 10^{10}\ \text{decays/s}, \tag{13.22}$$

matching the defined curie, 3.7×10^{10} Bq, to within rounding — a reminder that the "old" activity unit is not an arbitrary round number but a direct consequence of ^{226}Ra's actual half-life and molar mass, fixed once those two measured quantities are fixed.

Worked Example: Carbon-14 Dating

Living organisms continuously exchange carbon with the atmosphere, maintaining a constant fraction of the radioactive isotope $^{14}_{6}$C (half-life $T_{1/2} = 5730$ years, continuously replenished in the atmosphere by cosmic-ray-induced nuclear reactions) relative to stable ^{12}C; once an organism dies, that exchange stops, and the ^{14}C fraction — and hence the sample's ^{14}C activity — decays exponentially with no further replenishment. A wooden artifact is found to have a ^{14}C activity only 68.0% that of a freshly cut, otherwise identical sample of the same wood. Since activity is proportional to N at every instant, $\mathcal{A}/\mathcal{A}_0 = N/N_0 = 0.680$, and the decay law gives the artifact's age directly:

$$0.680 = e^{-\lambda t} \quad \Longrightarrow \quad t = -\frac{\ln(0.680)}{\lambda} = -\frac{\ln(0.680)}{\ln 2/T_{1/2}} = -\frac{(-0.386)(5730\ \text{yr})}{0.693} \approx 3.19 \times 10^{3}\ \text{yr}. \tag{13.23}$$

The artifact is roughly 3190 years old. This is exactly the calculation, worked from a real measurement rather than an assumed elapsed time, that Figure 13.4 above runs interactively — and it is also a direct illustration of why radiocarbon dating has a practical ceiling: after about ten half-lives ($\sim$ 57,000 years), the remaining ^{14}C activity has fallen by a factor of $2^{10} \approx 1000$, to a level low enough that statistical counting uncertainty in a realistic sample overwhelms the signal, and other methods (using much longer-lived parents, such as ^{238}U in the decay chain above) must take over.

Modes of Decay

Three principal decay modes connect unstable nuclides to more stable ones:

Alpha decay (${}^{A}_{Z}X \to {}^{A-4}_{Z-2}Y + \alpha$, where $\alpha = {}^{4}_{2}\mathrm{He}$) occurs predominantly among heavy nuclei, where it is energetically favorable (the parent's mass exceeds the combined daughter-plus-alpha mass) largely because of the Coulomb term discussed above. Classically, the alpha particle is confined within the nucleus by a potential well combining the short-range attractive strong force and, outside the nuclear radius, the repulsive Coulomb barrier — a barrier typically well above the alpha particle's actual kinetic energy once emitted, so classically the particle could never escape. Alpha decay is understood, quantitatively, as **quantum tunneling** (Chapter 8) of the alpha particle through this Coulomb barrier: the measured strong sensitivity of half-life to alpha particle energy (the **Geiger–Nuttall relation**, an empirical pattern in which small changes in emitted alpha energy correspond to enormous changes in half-life, spanning many orders of magnitude across known alpha emitters) is explained quantitatively by the exponential dependence of the tunneling probability on barrier height and width worked out in Chapter 8, making alpha decay one of the most direct large-scale confirmations of quantum tunneling. Concretely, for a heavy alpha emitter the Coulomb barrier peaks at a height of order 30–40 MeV (roughly the Coulomb repulsion of the departing alpha particle and daughter nucleus evaluated at the nuclear radius) while the emitted alpha particle typically carries away only 4–9 MeV of kinetic energy, so the barrier the alpha particle must tunnel through is tens of MeV high and, for a heavy emitter, typically several tens of femtometers wide between the nuclear surface and the outer classical turning point. The illustrative rectangular-barrier tunneling estimate worked out in Chapter 8 ($\Delta E \approx 20$ MeV, $L \approx 7$ fm, giving $T \approx e^{-2\kappa L}$ with $\kappa = \sqrt{2m_\alpha \Delta E}/\hbar$) is a deliberately simplified barrier with parameters of the same broad scale, not a literal width for every alpha emitter. Because T depends *exponentially* on both ΔE and L, a modest change in the alpha particle's energy routinely changes the predicted half-life by ten or more orders of magnitude — exactly the pattern captured empirically by the Geiger–Nuttall relation. Figure 13.7 watches that escape one nucleus at a time: the alpha particle's energy lies below the top of the Coulomb barrier, so it leaves by tunneling, each decay arrives at a random moment, and a sample of many nuclei settles onto the exponential the decay law predicts.

Beta decay occurs in three related forms — β^- decay ($n \to p + e^- + \bar{\nu}_e$, converting a neutron to a proton within the nucleus), β^+ decay ($p \to n + e^+ + \nu_e$), and electron capture ($p + e^- \to n + \nu_e$) — each mediated by the weak nuclear interaction (Chapter 14) and each moving a nucleus toward the more stable N/Z ratio for its mass number. The **neutrino** (ν_e) and **antineutrino** ($\bar{\nu}_e$) are required, not merely as bookkeeping devices, by conservation of energy, momentum, and angular momentum: without a third emitted particle, a two-body decay ($n \to p + e^-$ alone) would force the emitted electron to have one single, fixed energy for a given parent-daughter pair, but the observed electron energy spectrum in beta decay is continuous, spread over a range up to a fixed maximum — direct evidence (first argued by

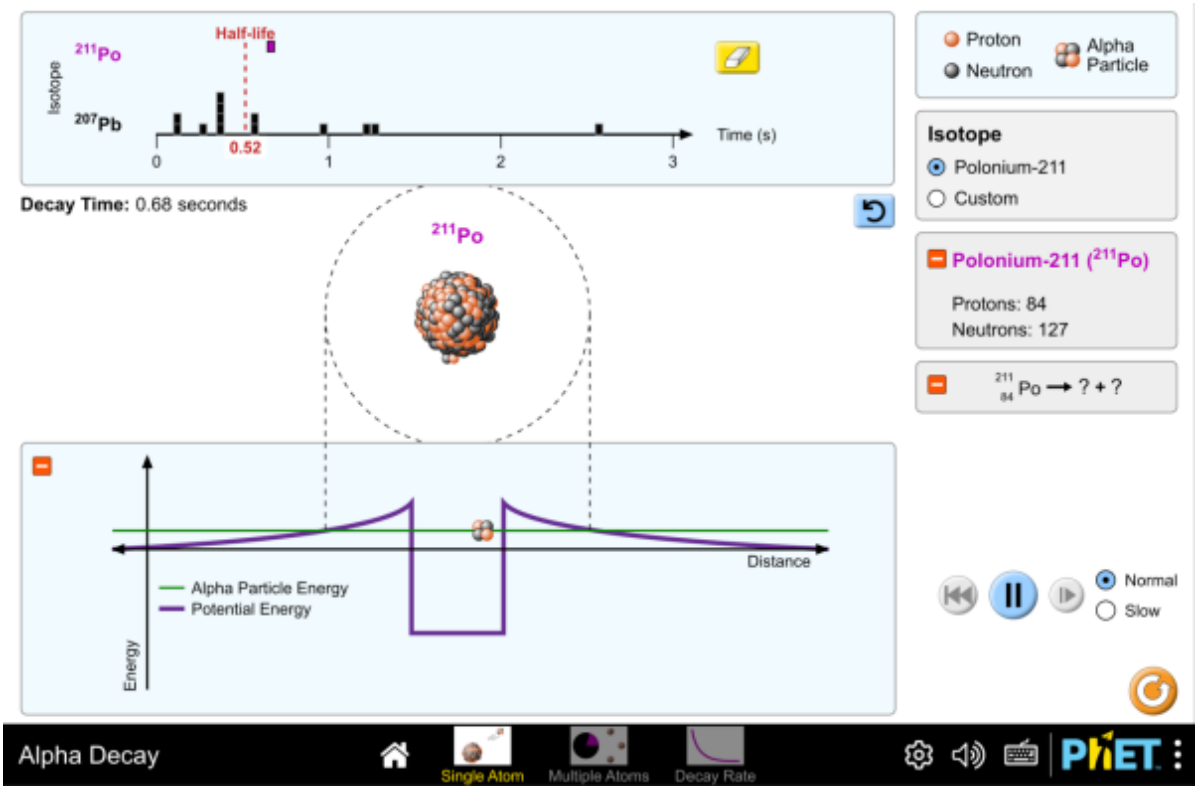

Figure 13.7: Alpha particles leaving a heavy nucleus by tunneling through the Coulomb barrier, and a sample of nuclei decaying with a measurable half-life. A single decay arrives at a random instant; a large sample follows the exponential of the decay law. Interactive simulation: Alpha Decay

Pauli in 1930, on exactly these grounds) that a third, initially unobserved particle carries away the missing energy and momentum event by event.

From Pauli's "Desperate Remedy" to a Confirmed Particle

Pauli's 1930 neutrino proposal — floated in an open letter to a nuclear physics conference he skipped, addressed to "Dear Radioactive Ladies and Gentlemen" — was, in his own words, a "desperate remedy" to rescue conservation of energy in beta decay, and he was privately doubtful it could ever be tested: the neutrino interacts so weakly that a typical one could pass through light-years of solid lead without interacting even once. It took until 1956, twenty-six years later, for Clyde Cowan and Frederick Reines to detect (anti)neutrinos directly, using the intense flux from a nuclear reactor and a target designed to catch the rare inverse-beta-decay event $\bar{\nu}_e + p \to n + e^+$ — a confirmation for which Reines shared the 1995 Nobel Prize in Physics (Cowan had died in 1974, and the Prize is never awarded posthumously).

The Weak Interaction, at the Quark Level

This chapter treats β^- decay at the nucleon level, $n \to p + e^- + \bar{\nu}_e$. Elementary Particles and the Standard Model goes one level deeper, showing the same transformation as a single down quark converting to an up quark — the only one of the four fundamental interactions that can change one flavor of quark into another — and develops the full conservation-law framework (charge, baryon number, lepton number) that the reaction above already silently obeys.

Gamma decay (${}^{A}_{Z}X^* \to {}^{A}_{Z}X + \gamma$, where the asterisk denotes an excited nuclear state) is the nuclear analog of atomic photon emission (Chapter 10): a nucleus left in an excited state, often as the immediate product of a preceding alpha or beta decay, drops to a lower-energy (often the ground) state by emitting a photon, with energy set by the spacing between nuclear energy levels — typically keV to MeV, far larger than atomic transition energies, because the nuclear scale of confinement is so much smaller than the atomic scale (an application of the same uncertainty-principle confinement argument used in Chapter 7).

All three modes, and the reason a given nuclide chooses one of them, are laid out in Figure 13.8. Its chart of the nuclides is the useful part: stable nuclides form a narrow band that starts along $N = Z$ and bends toward neutron excess as Z grows, and the decay mode of anything off that band is predictable from which side of it the nuclide sits on. Neutron-rich nuclides convert a neutron to a proton by β^- and step toward the band; proton-rich ones go the other way by β^+ or electron capture; and beyond $Z \approx 83$ nothing is stable at all, so the heavy corner of the chart empties itself by alpha emission.

Worked Example: The Alpha-Decay Q-Value for Uranium-238

Extending Exercise 13.4 into a full numerical calculation: for ${}^{238}_{92}\mathrm{U} \to {}^{234}_{90}\mathrm{Th} + \alpha$, the energy released (the **Q-value**) is the rest-mass energy of the parent minus that of the products, computed directly from tabulated atomic masses — no bare-proton-mass correction is needed here, since the same 92 electrons appear on both sides (90 bound to the thorium daughter, 2 bound to the emerging helium atom) and cancel exactly:

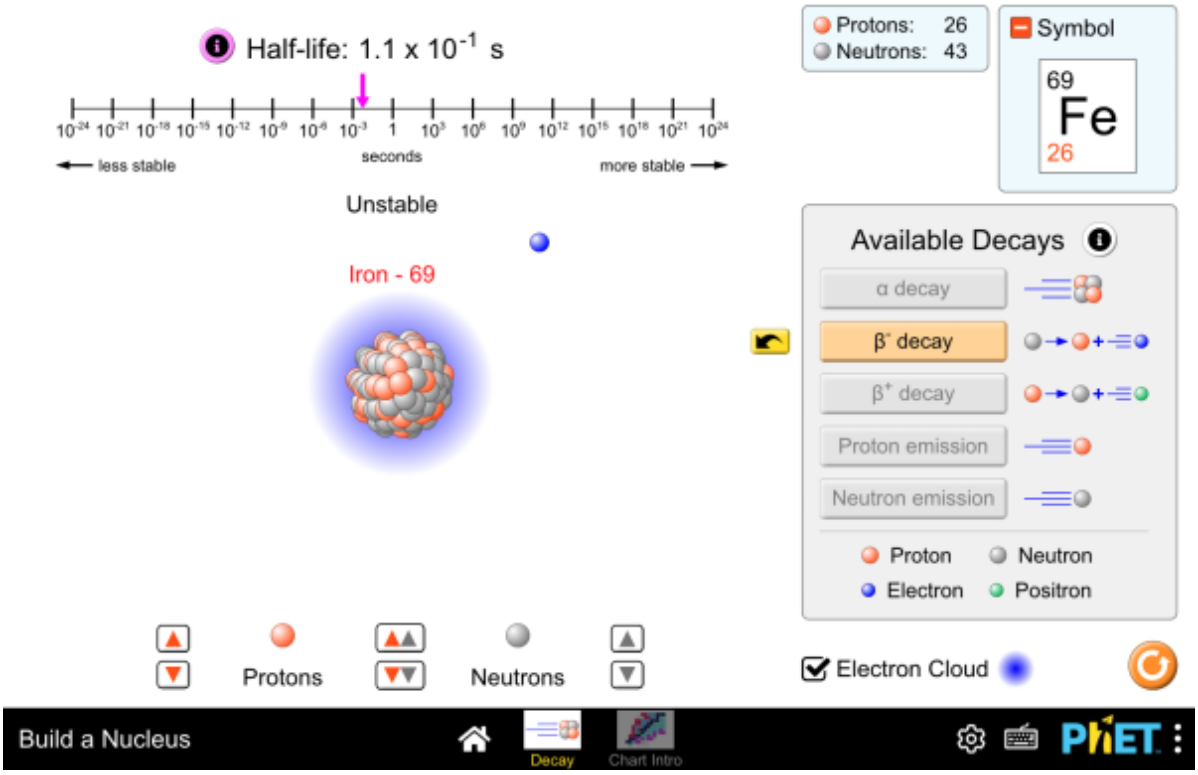

Figure 13.8: Nuclei assembled from protons and neutrons, with the decay mode of each unstable arrangement shown, alongside the chart of the nuclides. The chart is the whole of nuclear stability in one picture: a thin band of nuclides that last, and, on either side of it, a decay mode pointing back toward it. Interactive simulation: Build a Nucleus

$$Q = \left[M(^{238}\mathrm{U}) - M(^{234}\mathrm{Th}) - M(^{4}\mathrm{He})\right] c^2. \tag{13.24}$$

Using $M(^{238}\mathrm{U}) = 238.050788$ u, $M(^{234}\mathrm{Th}) = 234.043601$ u, and $M(^{4}\mathrm{He}) = 4.002602$ u,

$$Q = [238.050788 - 234.043601 - 4.002602]\,\mathrm{u}\times 931.494\,\mathrm{MeV/u} = (0.004585\,\mathrm{u})(931.494\,\mathrm{MeV/u}) \approx 4.27\,\mathrm{MeV}, \tag{13.25}$$

in close agreement with the measured value. Because momentum must also be conserved and the thorium daughter recoils, the alpha particle itself carries slightly less than the full Q: momentum conservation splits Q between the two products in inverse proportion to their masses, so the much lighter alpha particle ($m_\alpha \ll M_{\mathrm{Th}}$) carries away the large majority of it, close to the experimentally observed 4.20 MeV alpha kinetic energy for this decay, with the small remainder appearing as the recoiling thorium nucleus's kinetic energy.

Track Electrons Before Plugging In a Mass

Binding-energy and Q-value calculations mix "nuclear" bookkeeping (protons and neutrons) with tabulated *atomic* masses (nucleus plus electrons), and the two must be reconciled before any numbers go into a formula. The rule that keeps this honest: compare only mass values that carry the same total number of electrons. The ^{56}Fe binding-energy example above uses $m(^1\mathrm{H})$ in place of a bare m_p specifically so that all 26 atomic electrons cancel exactly rather than approximately; the ^{238}U alpha-decay Q-value works directly from atomic masses with no correction at all, because the same 92 electrons are present on both sides, split between the daughter thorium atom (90) and the emerging helium atom (2). Whenever the electron count does not match up this neatly, the mismatch has to be added back in by hand before the result can be trusted.

13.3 Fission and Fusion

Fission

Nuclear fission is the splitting of a heavy nucleus (typically after absorbing a neutron, which excites the nucleus into oscillation, distorting the initially spherical liquid drop) into two lighter, roughly comparable-mass fragments, plus several free neutrons. Because the binding-energy-per-nucleon curve rises steeply from heavy A toward the $A \approx 56$ peak, the fragments are individually more tightly bound (per nucleon) than the original heavy nucleus was, and the reaction releases a large amount of energy, typically around 200 MeV per fission event for a nucleus such as $^{235}_{92}\mathrm{U}$ — overwhelmingly larger than typical chemical reaction energies (electron-volts per bond, versus roughly 10^8 times more energy per fission event), directly reflecting the vastly greater strength of the nuclear force compared to the electromagnetic forces governing chemical bonding.

Because each fission event releases, on average, more than one free neutron, and each of those neutrons can potentially induce a further fission event in a neighboring nucleus, a **chain reaction** is possible if enough fissile material is present (a **critical mass**) to sustain, on average, at least one neutron-induced fission per neutron released — the basis of both controlled fission (nuclear power reactors, where the reaction rate is regulated, e.g. via neutron-absorbing control rods) and uncontrolled fission (fission weapons). A power reactor adds one further ingredient

beyond control rods: because the neutrons released promptly by fission are fast (born with roughly 1–2 MeV of kinetic energy) while ^{235}U fissions far more readily on absorbing a *slow* (thermal, ~ 0.025 eV) neutron, most reactor designs surround the fuel with a **moderator** — a light-nuclide material such as ordinary water, heavy water, or graphite — whose job is to slow fast neutrons toward thermal energies through repeated elastic collisions (most efficient, per collision, with nuclei of mass comparable to the neutron itself, which is why light nuclei make good moderators and heavy ones make poor ones) without absorbing them outright, so that they are far more likely to induce a further fission before escaping the reactor core or being captured non-productively.

The distinction between those last two outcomes is quantitative, not qualitative, and Figure 13.9 is where that becomes obvious. Fire a neutron at a single ^{235}U nucleus and watch the liquid drop distort, neck, and split; assemble a pile of nuclei instead and the multiplication factor decides everything. Below one neutron per fission surviving to cause another, the reaction dies out; above it, the population grows exponentially; and the control rods in the reactor screen do nothing more sophisticated than hold that number at exactly one by absorbing the surplus.

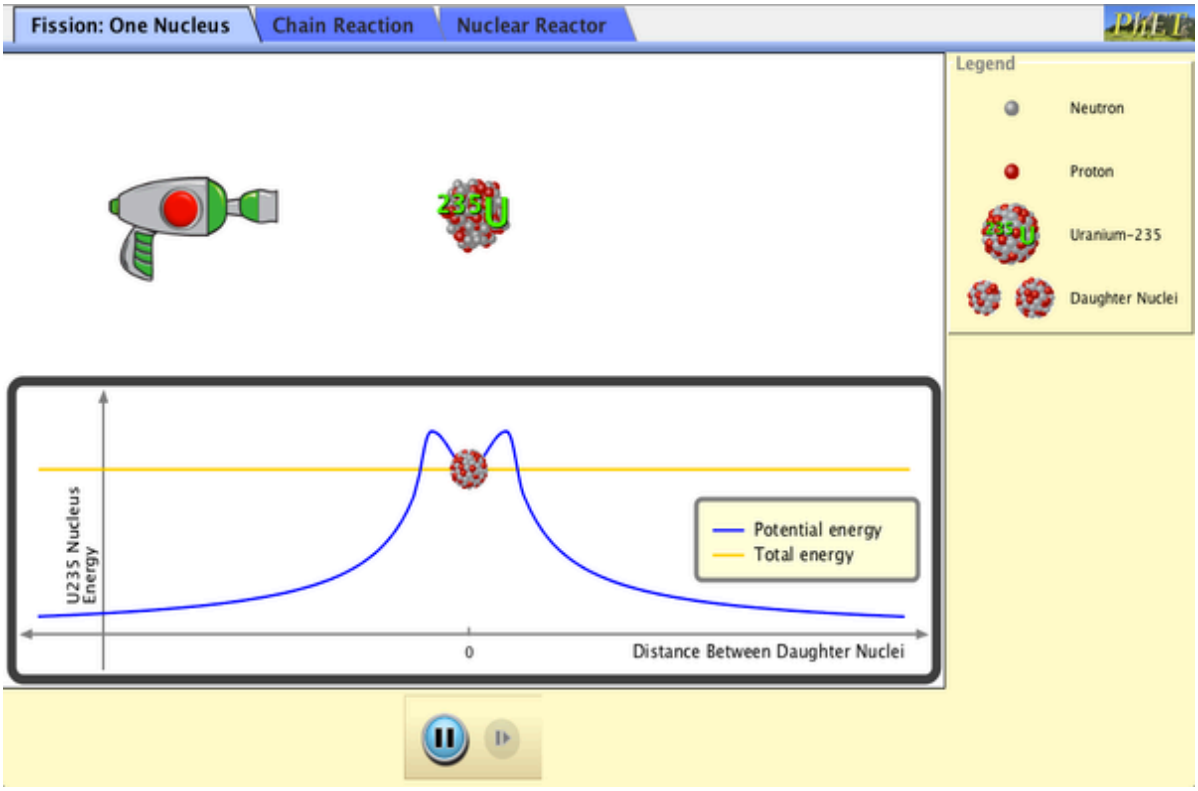

Figure 13.9: A single fission event, a chain reaction in an assembly of adjustable size and enrichment, and a controlled reactor. Criticality is a statement about a ratio, and it is the same ratio in all three. Interactive simulation: Nuclear Fission

Fusion

Nuclear fusion, the combination of two light nuclei into a single heavier one, releases energy for exactly the mirror-image reason: moving from very light A toward the peak of the binding-energy curve increases E_B/A, so the fused product is more tightly bound (per nucleon) than



Figure 13.10: A tour of a real reactor built around exactly the chain-reaction and moderator ideas developed above, from the control room down to the fuel assemblies. Watch on YouTube: I Explored the World's First Nuclear Power Plant (and How It Works)

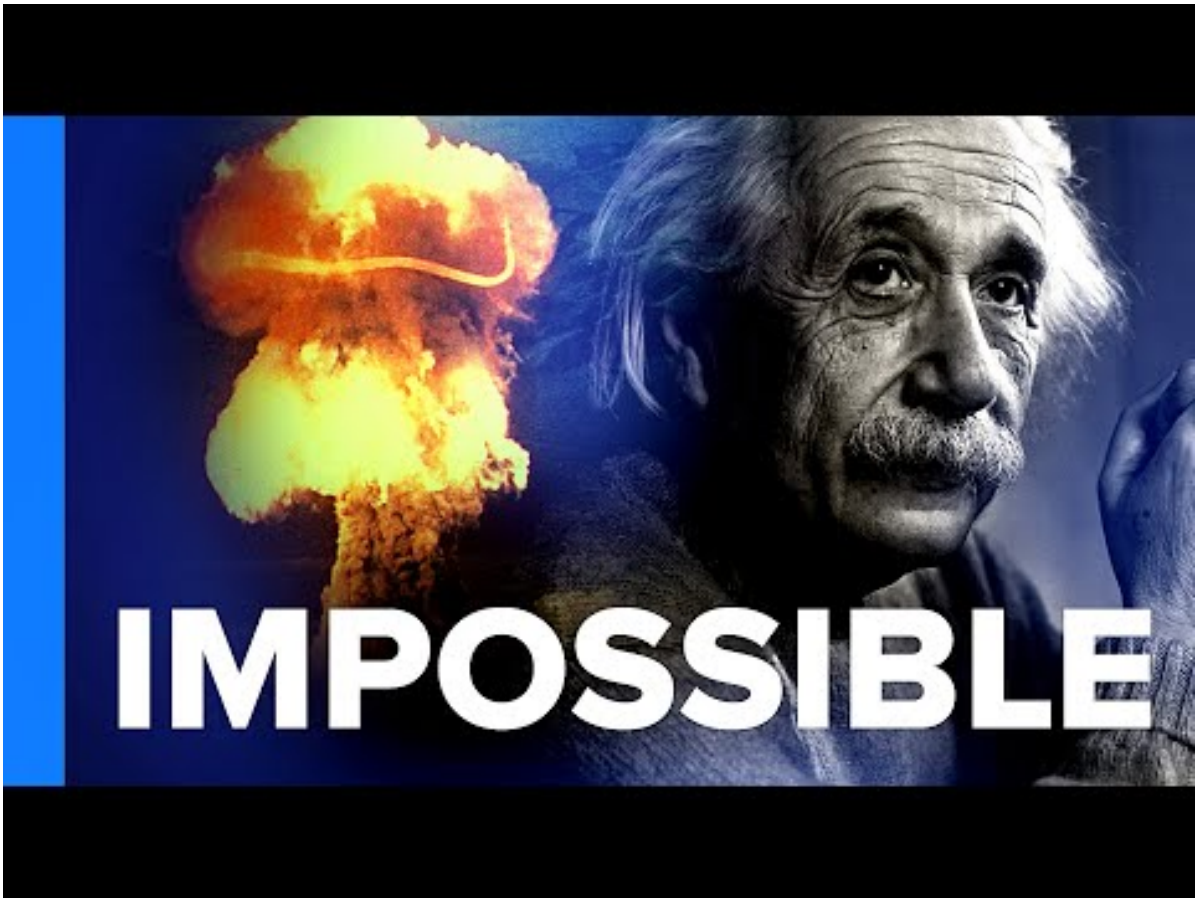

Figure 13.11: The same critical-mass reasoning from a different angle: why the numbers looked, at first, like they ruled out a chain-reaction weapon entirely. Watch on YouTube: Why Einstein Thought Nuclear Weapons Were Impossible

the separate light nuclei were. Fusion is the energy source that powers stars, where sequences of fusion reactions (in the Sun, predominantly the **proton-proton chain**, ultimately converting four protons into a helium-4 nucleus plus positrons, neutrinos, and gamma rays) release the energy that balances gravitational collapse and produces the Sun's luminosity (Chapter 3, Exercise 3.6). Because fusion requires two positively charged nuclei to approach to within the range of the strong force ($\sim$ 1 fm) against their mutual Coulomb repulsion, it requires very high temperatures (tens of millions of kelvin or more, as in stellar cores) to proceed at an appreciable rate even with the assistance of quantum tunneling through the Coulomb barrier — the same tunneling mechanism responsible for alpha decay, now working in reverse to allow two light nuclei to fuse despite insufficient classical kinetic energy to overcome their mutual repulsion.

The Proton-Proton Chain

Written out step by step, the dominant sequence in the Sun (the **proton-proton chain**, or pp I branch) is:

$$^{1}\mathrm{H} + {}^{1}\mathrm{H} \rightarrow {}^{2}\mathrm{H} + e^{+} + \nu_e \qquad \text{(one proton weak-converts to a neutron in the process)}, \tag{13.26}$$

$$^{2}\mathrm{H} + {}^{1}\mathrm{H} \rightarrow {}^{3}\mathrm{He} + \gamma, \tag{13.27}$$

$$^{3}\mathrm{He} + {}^{3}\mathrm{He} \rightarrow {}^{4}\mathrm{He} + 2\,{}^{1}\mathrm{H}, \tag{13.28}$$

with the first two steps each occurring twice to supply the two ^{3}He nuclei consumed in the third. Adding up all three steps and canceling the deuterium, helium-3, and one pair of protons that appear only as intermediates gives the net effect,

$$4\,{}^{1}\mathrm{H} \rightarrow {}^{4}\mathrm{He} + 2e^{+} + 2\nu_e, \tag{13.29}$$

converting four ordinary hydrogen nuclei into one helium-4 nucleus, two positrons (which promptly annihilate with ambient electrons, releasing further gamma-ray energy), and two neutrinos (which, interacting only weakly, escape the Sun's interior directly, carrying off a small fraction of the released energy without contributing to solar heating). The full sequence releases a net 26.7 MeV per helium-4 nucleus formed — about 0.7% of the rest-mass energy of the four consumed protons converted directly to energy — which, multiplied by the enormous number of such reactions occurring per second in the Sun's core, accounts for the Sun's entire luminosity (Chapter 3, Exercise 3.6). The first step — two protons fusing directly, with one simultaneously beta-plus-converting to a neutron — is both essential (it is the chain's only entry point from pure hydrogen) and extraordinarily slow, mediated by the weak interaction (Chapter 14) at the same instant as the strong-force capture, giving an individual proton in the Sun's core a mean waiting time of order *billions of years* before it fuses — the single biggest reason the Sun burns for billions rather than millions of years despite its core temperature and density being otherwise sufficient for fusion to proceed immediately.

Terrestrial fusion research targets a different, much faster reaction than the Sun's own, since the proton-proton chain's bottleneck first step is far too slow to be useful outside a star: **deuterium-tritium (D-T) fusion**,

$$^{2}_{1}\mathrm{H} + {}^{3}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + n, \tag{13.30}$$

pursued in magnetic-confinement **tokamak** reactors such as the ITER project in France — a fundamentally different confinement strategy from the Sun's, which needs no physical confinement at all beyond its own gravity; tokamaks instead use strong magnetic fields to hold a plasma far too hot ($\sim 10^8$ K, hotter than the Sun's core, needed to compensate for the much lower particle density achievable in a laboratory) for any physical vessel to contain directly.

Worked Example: Energy Released in Deuterium–Tritium Fusion

Using $m(^2\mathrm{H}) = 2.014102$ u, $m(^3\mathrm{H}) = 3.016049$ u, $m(^4\mathrm{He}) = 4.002602$ u, and $m_n = 1.008665$ u, the mass defect of the D-T reaction above is

$$\Delta m = [2.014102 + 3.016049] - [4.002602 + 1.008665] = 5.030151 - 5.011267 = 0.018884 \text{ u}, \tag{13.31}$$

so the energy released is

$$Q = \Delta m\, c^2 = (0.018884 \text{ u})(931.494 \text{ MeV/u}) \approx 17.6 \text{ MeV}, \tag{13.32}$$

in agreement with the accepted value. Conservation of momentum (the reacting nuclei have small initial momentum compared to the products' final momenta) splits this energy between the two products in inverse proportion to their masses, giving the alpha particle about 3.5 MeV and the neutron about 14.1 MeV of kinetic energy — a single D-T event releasing, in one step, roughly four times the energy of the U-238 alpha decay computed above, from a reaction between two of the lightest nuclei that exist rather than one of the heaviest.

13.4 Summary

- A nucleus ${}^{A}_{Z}X$ contains Z protons and $N = A - Z$ neutrons; nuclear radius scales as $R = R_0 A^{1/3}$, consistent with a roughly incompressible liquid-drop model of tightly packed nucleons.
- The short-range, charge-independent **strong nuclear force** overcomes Coulomb repulsion to bind the nucleus; **binding energy**, $E_B = [Zm_p + Nm_n - M]c^2$, and **binding energy per nucleon**, peaking near $A \approx 56$ (iron), summarize nuclear stability and are captured by the semi-empirical mass formula's volume, surface, Coulomb, symmetry, and pairing terms.
- Radioactive decay follows the exponential law $N(t) = N_0 e^{-\lambda t}$, with half-life $T_{1/2} = \ln 2/\lambda$ and activity $\mathcal{A} = \lambda N$.
- **Alpha decay** proceeds by quantum tunneling through the Coulomb barrier; **beta decay** (β^-, β^+, electron capture) proceeds via the weak interaction and requires a neutrino/antineutrino to conserve energy, momentum, and angular momentum, as shown by the continuous beta-electron energy spectrum; **gamma decay** is photon emission between nuclear energy levels.
- **Fission** (heavy nucleus splits) and **fusion** (light nuclei combine) both release energy by moving nucleons toward the binding-energy peak near $A \approx 56$; fission chain reactions require a critical mass and, in a reactor, a **moderator** to slow fission-born fast neutrons, and fusion requires overcoming the Coulomb barrier, typically via high temperature and tunneling, as in stellar interiors.
- The **semi-empirical (Weizsäcker) mass formula** writes E_B as volume ($a_V A$), surface ($-a_S A^{2/3}$), Coulomb ($-a_C Z(Z-1)A^{-1/3}$), symmetry ($-a_A(A-2Z)^2/A$), and pairing

($\pm a_P A^{-3/4}$, or 0) terms; the pairing term explains why even-even nuclides dominate the stable chart (and stable odd-odd nuclides are essentially absent), and maximizing E_B at fixed A shows why heavy stable nuclei favor $N > Z$. The **nuclear shell model** (an average central potential plus spin-orbit coupling) explains the extra stability of nuclides at the magic numbers $2, 8, 20, 28, 50, 82, 126$, the nuclear analog of noble-gas electron-shell closure.

- A radioactive **decay chain** links successive unstable nuclides (e.g. ^{238}U through fourteen alpha and beta decays to stable ^{206}Pb) until a stable nuclide is reached, reaching **secular equilibrium** when run undisturbed for long enough; the same exponential decay law underlies **radiocarbon dating** of once-living material.
- **Fusion** proceeds in stars via the multi-step **proton-proton chain** ($4\,^1\mathrm{H} \to {}^4\mathrm{He} + 2e^+ + 2\nu_e$, net 26.7 MeV) and in terrestrial reactors via the faster **deuterium-tritium reaction** (17.6 MeV per event), the latter pursued in magnetic-confinement devices such as tokamaks.

13.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens on the binding-energy curve.

1. **Multiple choice, from a figure.** Why can both fusion of light nuclei and fission of heavy nuclei release energy? Figure description: Binding energy per nucleon versus mass number, peaking near iron and nickel. Choices: (A) Both move the products toward higher binding energy per nucleon near iron and nickel. (B) Both convert every nucleon entirely into radiation. (C) Both eliminate the strong nuclear interaction. (D) Both always increase the total rest mass of the products.

2. **True or false.** A radioactive nucleus becomes more likely to decay merely because it has survived for a long time.

3. **Drag the words.** Complete the nuclear-decay descriptions. Alpha decay proceeds by _____. Beta decay uses the _____. Gamma decay changes nuclear _____. Radioactive populations follow an _____. (Terms: quantum tunneling; weak interaction; energy without changing A or Z; exponential law; electromagnetic tunneling of electrons; linear law.)

4. **Fill in the blanks.** Complete the nuclear-size and decay relations. Nuclear radius scales as A raised to the power _____, and the half-life equals ln 2 divided by the decay _____.

5. **Mark the words.** Mark the two processes that move nuclei toward the binding-energy peak. “Energy can be released by fusion of light nuclei and by fission of very heavy nuclei. Both produce more tightly bound products.”

13.6 Conceptual Questions

1. Why does the short range of the strong force lead to approximately constant nuclear density?

2. Distinguish mass defect from binding energy while explaining why they contain the same physical information.

3. Why is radioactive decay exponential even though no nucleus becomes more likely to decay merely because it is old?

4. How do alpha, beta, and gamma decay change A and Z differently?

5. Why are slow neutrons especially effective at sustaining fission in many reactors?

6. Fusion releases energy, yet positively charged nuclei repel. What physical barrier must a reactor or star overcome?

13.7 Problems

Exercise 13.1
(Straightforward)

Estimate the nuclear radius of $^{238}_{92}\mathrm{U}$ and of $^{4}_{2}\mathrm{He}$ using $R = R_0 A^{1/3}$, and compute the ratio of their radii. Comment on whether this ratio is consistent with $A^{1/3}$ scaling given the ratio of their mass numbers.

Exercise 13.2
(Straightforward)

Compute the binding energy and binding energy per nucleon of $^{4}_{2}\mathrm{He}$, given $M(^{4}_{2}\mathrm{He}) = 4.002602$ u, $m_p = 1.007276$ u, $m_n = 1.008665$ u, and using $1\ \mathrm{u}\,c^2 = 931.5$ MeV (you may neglect the small correction from atomic electron binding energies).

Exercise 13.3
(Straightforward)

A radioactive sample of $^{131}_{53}\mathrm{I}$ (half-life 8.02 days) initially contains $N_0 = 1.00 \times 10^{18}$ nuclei. (a) Find the decay constant λ. (b) Find the number of nuclei remaining after 24 days. (c) Find the initial activity, in becquerels (decays/s).

Exercise 13.4
(Straightforward)

$^{238}_{92}\mathrm{U}$ undergoes alpha decay to $^{234}_{90}\mathrm{Th}$. Write the full decay equation, and explain, using the shape of the binding-energy-per-nucleon curve, why alpha decay (rather than, say, single-proton emission) is the energetically favored decay mode for very heavy nuclei.

Exercise 13.5
(Moderate)

In beta-minus decay of a free neutron, $n \to p + e^- + \bar{\nu}_e$, use $m_n c^2 = 939.57$ MeV, $m_p c^2 = 938.27$ MeV, $m_e c^2 = 0.511$ MeV (and treat the antineutrino as massless) to find the total kinetic energy shared among the three decay products. Explain why the electron's kinetic energy alone is not fixed at this value, but instead varies continuously up to it.

Exercise 13.6
(Moderate)

In the fission of $^{235}_{92}\mathrm{U}$ (via neutron absorption to $^{236}_{92}\mathrm{U}$, then fission), roughly 200 MeV is released per event. Estimate the mass (in kg) of $^{235}\mathrm{U}$ that would need to fission completely to release 1.0×10^{14} J (order of magnitude of a small commercial reactor's daily output), using Avogadro's number and the molar mass of $^{235}\mathrm{U}$.

Exercise 13.7
(Moderate)

Using the semi-empirical mass formula coefficients given in the text, compute the predicted binding energy per nucleon of $^{120}_{50}\mathrm{Sn}$ ($N = 70$, even–even). Compare your result to the measured value, 8.51 MeV/nucleon, and comment on the size and likely origin of any discrepancy.

Exercise 13.8
(Moderate)

In the uranium-238 decay chain, $^{226}_{88}\mathrm{Ra}$ (half-life 1600 yr) alpha-decays to $^{222}_{86}\mathrm{Rn}$ (half-life 3.8 days). A sealed sample containing 1.00 g of $^{226}\mathrm{Ra}$ has been undisturbed for many thousands of years — far longer than radon's half-life, so the radon has reached secular equilibrium with its radium parent. Using the result of the radium worked example, find the activity, in becquerels, of the $^{222}\mathrm{Rn}$ in the sample. (Hint: what must the daughter's activity equal, in secular equilibrium?)

Exercise 13.9
(Challenging)

A bone fragment has a measured $^{14}\mathrm{C}$ activity that is 42.0% of that of living bone. Using $T_{1/2} = 5730$ years for $^{14}\mathrm{C}$, find the fragment's age.

Exercise 13.10
(Challenging)

Verify the claim in the deuterium-tritium fusion worked example that momentum conservation splits the reaction's 17.6 MeV Q-value into 3.5 MeV (carried by the alpha particle) and 14.1 MeV (carried by the neutron): treating the initial deuteron and triton as essentially at rest, set the magnitudes of the final alpha-particle and neutron momenta equal (nonrelativistic momentum conservation) and solve for each product's share of the total kinetic energy in terms of the two product masses ($m_\alpha \approx 4.00$ u, $m_n \approx 1.01$ u).

Chapter 14

Elementary Particles and the Standard Model

Learning Objectives

By the end of this chapter, you should be able to:

- Classify particles as fermions or bosons and explain the physical distinction.
- Distinguish the four fundamental interactions by their relative strength, range, and mediating boson.
- Distinguish leptons and quarks as the two families of fundamental fermions, and state the generation structure of each.
- Explain quark confinement and construct the quark content of simple baryons and mesons.
- Apply conservation laws (charge, baryon number, lepton number, and strangeness) to determine whether a proposed particle reaction is allowed, and identify which interaction (strong, electromagnetic, or weak) a given reaction must proceed through.
- Describe the role of the Higgs field in giving mass to fundamental particles.
- Read and interpret a simple Feynman diagram, including the convention that an antiparticle's arrow points backward in time along a line that itself still runs forward.
- Describe, in outline, how particle accelerators and detectors produce and identify new particles, and explain why colliders are favored over fixed-target machines for reaching the highest energies.
- Recount the historical "particle zoo" of the 1930s–1960s and explain how the eightfold way and the quark model resolved it.
- Estimate, quantitatively, why gravitational effects between individual particles are negligible compared with the other fundamental interactions.
- Identify several of the major open questions at the current frontier of particle physics and cosmology.

Introduction

Nuclear physics (Chapter 13) treated protons and neutrons as elementary building blocks, bound by the strong force into nuclei. This chapter goes one level deeper, to the particles and forces from which protons, neutrons, and indeed all matter are built. Over the course of the twentieth century, a rapidly growing zoo of "elementary" particles, discovered first in cosmic rays and then in purpose-built particle accelerators, was gradually organized into a

coherent theoretical framework, the **Standard Model of particle physics**, which identifies a genuinely small set of truly fundamental particles and the forces (themselves mediated by particle exchange) that govern their interactions. This chapter surveys that framework: the classification of particles by spin and by the forces they feel, the substructure of protons and neutrons in terms of quarks, the conservation laws that govern which particle reactions can occur, and the mechanism by which most fundamental particles acquire mass at all. It also traces how this framework was actually established — the historical progression from a bewildering "particle zoo" of cosmic-ray discoveries to the ordered quark model, and the particle accelerators and detectors that turned particle physics into a precision experimental science — and closes, as befits the final chapter of this book, with several of the major questions the Standard Model leaves unanswered.

14.1 Particles, Forces, and Feynman Diagrams

Classifying Particles: Fermions and Bosons

Every known particle carries an intrinsic spin angular momentum, quantized exactly as orbital angular momentum is (Chapter 9), $S = \sqrt{s(s+1)}\hbar$, but with s either a half-integer $(\frac{1}{2}, \frac{3}{2}, \ldots)$ or an integer $(0, 1, 2, \ldots)$. This distinction is far more than bookkeeping: particles with half-integer spin, called **fermions**, obey the Pauli exclusion principle (Chapter 11) and cannot occupy the same quantum state as an identical partner; particles with integer spin, called **bosons**, obey no such restriction and can occupy the same state in unlimited numbers (the basis, for photons, of the stimulated-emission cascade in a laser, Chapter 11). Matter, in the Standard Model, is built from fermions (spin $\frac{1}{2}$); the forces between them are mediated by the exchange of bosons.

> The names honor statistics, not the particles themselves: **fermions** obey Fermi–Dirac statistics, after Enrico Fermi, while **bosons** obey Bose–Einstein statistics, after Satyendra Nath Bose, whose 1924 analysis of photon statistics (soon extended by Einstein to massive particles) first showed that indistinguishable integer-spin particles behave nothing like classical particles, or like half-integer-spin fermions.

The Four Fundamental Interactions

All observed particle interactions are, to the precision of current experiments, accounted for by exactly four fundamental forces, each mediated by the exchange of a characteristic boson (a **gauge boson**, in the language of the quantum field theories underlying the Standard Model) and each with a distinct characteristic strength and range:

The interactions are compared directly in Table 14.1.

The weak interaction and the residual strong interaction between nucleons are confined to nuclear-scale distances, which is why their effects were unknown until nuclear and particle physics probed those scales directly. The underlying strong interaction is instead a confining interaction between quarks; the electromagnetic and gravitational interactions are infinite in range because their mediating bosons are massless, and they dominate everyday, macroscopic experience. The relative strengths quoted are approximate and depend on the energy/distance scale at which the comparison is made, but the ordering — strong $\gg$ electromagnetic $\gg$ weak $\gg$ gravitational, at typical particle-physics scales — is robust. The strong force between

Table 14.1: The four fundamental interactions

Interaction	Relative strength	Range	Mediating boson(s)	Governs
Strong	1	confining; residual range ~ 1 fm	gluon (QCD); mesons for the residual force	quark binding; nuclear force (residual)
Electromagnetic	$\sim 10^{-2}$	infinite	photon	charged-particle interactions
Weak	$\sim 10^{-6}$	$\sim 10^{-3}$ fm	$W^{\pm}$, Z^0	beta decay; some particle decays
Gravitational	$\sim 10^{-38}$	infinite	graviton (hypothesized, not yet observed)	negligible for individual particles; dominant only for macroscopic masses

nucleons discussed in Chapter 13 is, in the Standard Model's deeper description, a residual effect of the strong force acting between the quarks confined inside each nucleon (analogous to the way the residual electromagnetic force between neutral atoms, the van der Waals force, is a residual effect of the more fundamental electromagnetic force between the charged constituents of each atom).

Worked Example: Gravity Versus Electromagnetism for Two Protons

Table 14.1 states that gravity is weaker than the strong interaction by a factor of order 10^{-38}; it is worth seeing this concretely for the simplest possible case — two protons separated by some distance r — where both the gravitational attraction and the electric (Coulomb) repulsion are exactly known, textbook formulas requiring no particle-physics machinery at all:

$$\frac{F_{\text{grav}}}{F_{\text{EM}}} = \frac{Gm_p^2/r^2}{ke^2/r^2} = \frac{Gm_p^2}{ke^2}, \tag{14.1}$$

independent of r, since both forces obey the same inverse-square law and the separation cancels out of the ratio. Substituting $G = 6.674 \times 10^{-11}\ \text{N}\cdot\text{m}^2/\text{kg}^2$, $m_p = 1.673 \times 10^{-27}$ kg, $k = 8.988 \times 10^{9}\ \text{N}\cdot\text{m}^2/\text{C}^2$, and $e = 1.602 \times 10^{-19}$ C,

$$\frac{F_{\text{grav}}}{F_{\text{EM}}} = \frac{(6.674 \times 10^{-11})(1.673 \times 10^{-27})^2}{(8.988 \times 10^{9})(1.602 \times 10^{-19})^2} \approx 8 \times 10^{-37}, \tag{14.2}$$

consistent, to within the precision to which either figure is normally quoted, with the ratio of the two relative-strength entries in Table 14.1 ($10^{-38}/10^{-2} = 10^{-36}$). Gravity between two individual protons is roughly 36 orders of magnitude weaker than their mutual electric repulsion — so extreme that no experiment has ever measured the gravitational force between two elementary particles directly; it is simply swamped by every other force present. Gravity dominates only on macroscopic and astronomical scales because mass is exclusively attractive and accumulates without limit (every proton in a star adds to the same gravitational pull), whereas electric charge comes in both signs and ordinary bulk matter is almost perfectly charge-neutral, so the electromagnetic forces of its enormous number of individual charges cancel to

a degree gravity's never does. This is also, precisely, why the graviton remains hypothetical in Table 14.1: a force this feeble, acting between single particles, is far beyond the reach of any conceivable particle-physics detector, and quantum gravity remains an unsolved theoretical problem rather than an experimentally accessible one, as the Open Questions section at the end of this chapter discusses further.

Of the four, the weak interaction is the one with no everyday signature to point at — it is too short-ranged to hold anything together and too feeble to push anything around — and its importance is easy to underrate as a result. Its distinction is that it is the only interaction that changes one kind of quark into another, and Figure 14.1 is that transformation at work: a neutron (udd) becomes a proton (uud) by turning a down quark into an up quark, with an electron and an antineutrino carrying off the charge and lepton-number balance. None of the other interactions in Table 14.1 can do this. Without it, the proton-proton chain could never take its first step, no hydrogen would ever become helium, and stars would not shine.

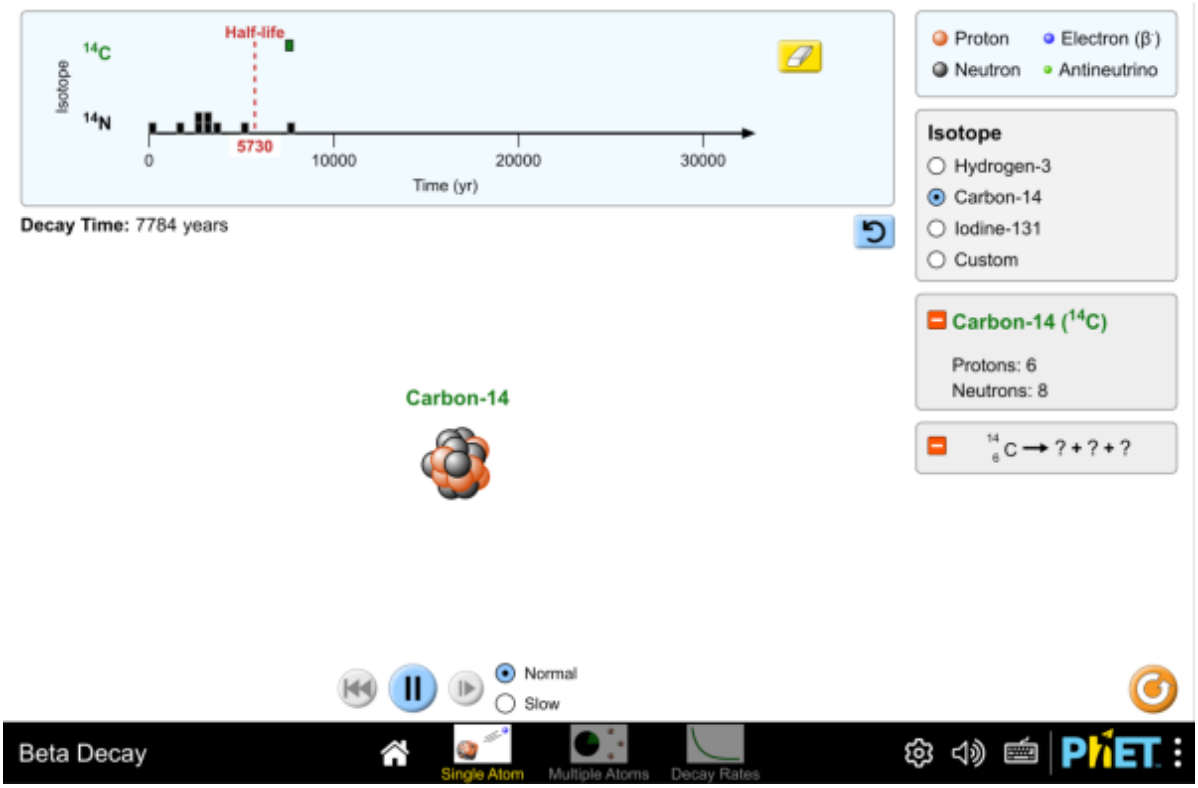

Figure 14.1: Beta decay of a single nucleus and of a sample, with the emitted electron and antineutrino shown. The conservation laws cataloged later in this chapter can be read directly off the products: charge, baryon number, and electron lepton number each balance, event by event. Interactive simulation: Beta Decay

Reading Feynman Diagrams

Every interaction cataloged in Table 14.1 — one particle scattering off another, one particle decaying into several — can be drawn as a **Feynman diagram**, a bookkeeping and visualization tool introduced by Richard Feynman in the late 1940s as part of the machinery that made precise calculations in quantum electrodynamics tractable (work for which Feynman shared the 1965 Nobel Prize with Julian Schwinger and Sin-Itiro Tomonaga). A Feynman diagram is

not a literal picture of particle trajectories in space, the way a cloud-chamber photograph is; it is closer to a schematic circuit diagram, read according to a small set of fixed conventions:

- The vertical axis is **time**, increasing upward (some texts instead run time left to right; the convention adopted here matches the light-cone diagrams of Chapter 2). The horizontal axis is space, drawn schematically and not to scale.
- A straight line represents a **fermion** (a matter particle), with an arrow indicating the direction of particle flow. A particle's arrow points forward in time, in the direction the line is actually traversed; an **antiparticle** is drawn as its particle's line traversed backward — the arrow points backward in time even though, like everything else in the diagram, the antiparticle itself only ever moves forward in time.
- A wavy or dashed internal line represents the **exchanged boson** that mediates the interaction — a photon, gluon, or $W^{\pm}/Z^0$, depending on which force is at work.
- A **vertex**, where lines meet, represents a single interaction. Every exactly-conserved quantity introduced in this chapter — electric charge above all — must balance at *each individual vertex*, not merely for the diagram as a whole, since a vertex is the fundamental, irreducible unit of interaction the Standard Model actually specifies.

The internal boson line is called **virtual**: unlike an external line (an actual incoming or outgoing particle, which must satisfy the energy–momentum relation $E^2 = (pc)^2 + (mc^2)^2$ of Chapter 3 for its own rest mass), a virtual particle is an internal quantum-mechanical contribution and need not satisfy that relation (it is not an independently observable particle). The often-used language of "borrowing" energy from the energy–time uncertainty principle (Chapter 7) is only a heuristic; energy and momentum remain conserved at every vertex. The finite range associated with exchange of a massive boson can be estimated from its Compton wavelength, $\hbar/(mc)$: the massive $W^{\pm}$ and Z^0 ($m_W c^2 \approx 80.4$ GeV) confine the weak interaction to a range of only $\hbar c/(m_W c^2) \approx (197\ \text{MeV}\cdot\text{fm})/(8.04\times10^4\ \text{MeV}) \approx 2\times10^{-3}$ fm — matching the table's entry to within the precision to which it is normally quoted. The photon is massless and mediates an infinite-range force. The residual strong force between nucleons is instead finite-ranged, dominated at ordinary nuclear distances by exchange of massive mesons such as pions; the gluons themselves are confined and do not produce a freely propagating long-range force between nucleons.

Figure 14.2 shows two of the simplest possible diagrams, one for each of two very different interactions.

The left-hand diagram is the quantum-field-theoretic picture underlying ordinary Coulomb repulsion: two electrons never touch, but exchange a virtual photon at a single vertex on each of their two lines, each vertex conserving charge trivially (an electron's charge is unchanged by absorbing or emitting an uncharged photon). The right-hand diagram redraws the beta-decay process already introduced above, but at the level of an individual quark rather than the nucleon as a whole: it is really the neutron's down quark that emits the virtual W^- and converts directly to an up quark, turning udd into uud, while the neutron's other two quarks look on as unaffected spectators. This is also the precise reason the weak interaction, uniquely among the four, can change a quark's flavor at all. At a photon or gluon vertex, the boson emitted or absorbed carries zero electric charge, so charge conservation alone forces the quark's charge — and hence its flavor family — to be exactly the same before and after; a d quark ($-\frac{1}{3}e$) simply cannot become a u quark ($+\frac{2}{3}e$) at such a vertex without violating charge conservation right there at the vertex. The $W^{\pm}$ boson is the only carrier with nonzero electric charge ($\pm e$), and $\pm e$ is exactly the charge difference between an up-type and a down-type quark in the same generation ($Q(u) - Q(d) = \frac{2}{3}e - (-\frac{1}{3}e) = e$), so only a $W^{\pm}$ vertex has the right charge budget

Reading a Feynman diagram: time runs upward, a vertex is an interaction, an internal wavy line is the exchanged (virtual) boson

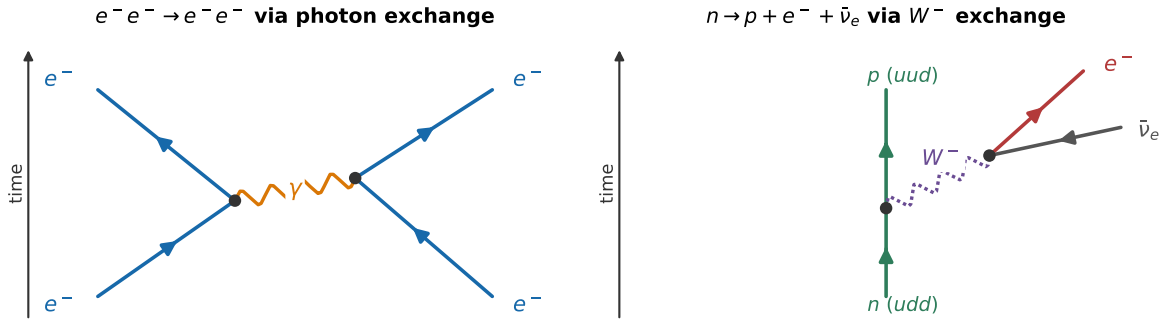

Figure 14.2: Left: two electrons scatter by exchanging a virtual photon — the same photon whose particle nature was established in Chapter 6, here playing the role of the electromagnetic force carrier at the vertex level. Right: a neutron decays to a proton by emitting a virtual W^-, which itself decays to an electron and an electron antineutrino; the antineutrino's arrow points backward in time, marking it as an antiparticle. Original schematic by the author.

to convert one into the other.

> **Warning**
> Do not read the backward arrow on an antiparticle line as motion backward in time. The line still advances forward through the diagram; the arrow is a bookkeeping convention for particle flow and conserved charge.

The generic term for a quark or lepton type — up, down, strange, charm, top, bottom for quarks; electron, muon, tau (and their neutrinos) for leptons — is **flavor**. Flavor is independent of the color charge introduced later in this chapter: two quarks of the same flavor, two up quarks say, can still carry different colors.

Worked Example: Reading a Feynman Diagram

Verify, vertex by vertex, that the beta-decay diagram in Figure 14.2 conserves charge, baryon number, and lepton number at *each* vertex individually, not merely for the overall reaction $n \to p + e^- + \bar{\nu}_e$ already checked earlier in this chapter.

First vertex (shown schematically at the nucleon level as $n \to p + W^-$): charge before is 0 (the neutron); charge after is $Q(p) + Q(W^-) = (+e) + (-e) = 0$. Balanced. Baryon number before is $+1$ (n); after, $B(p) = +1$ and $B(W^-) = 0$ (the W is a boson, carrying no baryon number), so $B_{\text{after}} = +1$. Balanced. At the underlying quark vertex this is $d \to u + W^-$, with the neutron's other two quarks acting as spectators.

Second vertex ($W^- \to e^- + \bar{\nu}_e$): charge before is $-e$ (the W^-); charge after is $Q(e^-) + Q(\bar{\nu}_e) = (-e) + 0 = -e$. Balanced. Lepton number (L_e) before is 0 (the W carries none); after, $L_e(e^-) = +1$ and $L_e(\bar{\nu}_e) = -1$, summing to 0. Balanced.

Every conservation law holds separately at each vertex, which is both a check on the diagram and the deeper reason the *overall* reaction conserves these quantities in the first place: a multi-vertex process can only conserve a quantity globally if every individual vertex conserves it, since nothing is allowed to leak in or out of an internal line that connects two vertices back to back.

14.2 Accelerators, Leptons, and Quarks

Particle Accelerators and Detectors

None of the particle content, forces, or conservation laws cataloged in this chapter were read directly off nature; they were inferred, often at great technical and financial cost, from what came out of high-energy collisions. A **particle accelerator** uses electric fields to increase a charged particle's kinetic energy, typically bending its path into a circle with magnetic fields (a **synchrotron**, of exactly the type used at the Bevatron introduced in Chapter 3) so that the same beam can be accelerated repeatedly on each pass rather than only once, as in a straight-line (linear) accelerator. That beam is then steered into a collision — either against a stationary target (**fixed-target**) or head-on against a second, oppositely circulating beam (a **collider**) — and the CM-frame threshold-energy analysis of Chapter 3 governs, in either case, exactly how much beam energy is needed to produce a given final state.

What comes out of the collision is reconstructed by a **detector**, typically built in concentric layers immediately surrounding the collision point: an inner **tracking** layer, threaded by a magnetic field, records the curved paths of charged particles, whose momentum is read off from the radius of curvature via $p = qBr$ — the same relation used to test $p = \gamma mu$ directly in Chapter 3; beyond that, a **calorimeter** stops most particles outright and measures their energy from the size of the resulting shower of secondary particles; and an outermost layer specifically identifies **muons**, since a muon, carrying no strong-interaction charge and losing energy only slowly, is typically the only charged particle able to punch all the way through the calorimeter's absorbing material. Combining the tracking, calorimetry, and timing information from every layer, for millions of collisions per second, is how a modern detector reconstructs which particles were actually produced in each individual event — the only way any of the particles, decay modes, and conservation laws in this chapter were ever actually established.

The largest such machine ever built is the Large Hadron Collider (**LHC**), operated by CERN outside Geneva: a 27-kilometer circular tunnel in which two counter-circulating proton beams are accelerated to several TeV each (1 TeV $= 10^3$ GeV $= 10^6$ MeV) before colliding head-on at four points around the ring, each instrumented with a detector of exactly the layered design just described. Two of those detectors, ATLAS and CMS, are shown mid-construction in Figure 14.3; it was data from detectors of this kind, sifted from many billions of recorded collisions, that produced the statistically decisive evidence for the Higgs boson announced in 2012 and discussed further below.

> Particle physicists routinely quote a particle's rest mass directly in energy units via $E = mc^2$, dropping the "$/c^2$" as understood: "the W boson's mass is 80.4 GeV" and "$m_W c^2 = 80.4$ GeV," used earlier in this chapter, mean exactly the same thing. Taken to its logical conclusion, this is the field's **natural-unit** convention $c = \hbar = 1$, in which mass, momentum, and energy all share a single unit.



Figure 14.3: The ATLAS detector at CERN, photographed during construction in February 2007; the detector's wheel-shaped toroid magnets, used to bend and measure the momentum of outgoing muons, are visible at either side. Photograph by Sindre Skrede, 2007; released into the public domain via Wikimedia Commons.

Not every particle produced in a collision leaves a readily visible signal. A neutrino — carrying no electric charge and feeling only the weak interaction — usually passes through even a large, modern detector's tracking, calorimetry, and muon layers essentially undisturbed, exactly as it passes through the Earth itself. A small fraction do undergo a weak interaction in detector material, and those interactions provide direct event-by-event neutrino detections in dedicated neutrino experiments. In a collider detector, however, an escaping neutrino is usually inferred from missing transverse momentum, much as Pauli inferred the neutrino's existence from the continuous beta-decay energy spectrum (Chapter 13), but applied now to *momentum*: because the incoming beams have negligible net transverse momentum, any transverse momentum imbalance among everything the detector *does* see may have been carried away by one or more unseen particles. Detector resolution and unobserved beam remnants can also contribute to the imbalance, so missing transverse momentum is evidence rather than an exact identification by itself.

Self-check

If a neutrino escapes unseen, which measured quantity tells you something is missing? Why do head-on colliders reach higher center-of-momentum energy than fixed-target machines for the same beam energy?

Worked Example: Inferring a Missing Neutrino from Momentum Conservation

Suppose a collision event at a detector produces several visible particles (tracked and measured by the calorimeter) whose momentum components transverse to the beam sum to $p_x = 12.4$ GeV/c and $p_y = -5.1$ GeV/c. Since the incoming beams carry zero net transverse momentum by construction, conservation of momentum requires the missing transverse momentum, carried by one or more undetected particles (a neutrino, most often), to be exactly the negative of the visible sum:

$$p_x^{\text{miss}} = -12.4 \text{ GeV}/c, \qquad p_y^{\text{miss}} = +5.1 \text{ GeV}/c, \tag{14.3}$$

with magnitude $|\vec{p}^{\,\text{miss}}| = \sqrt{(12.4)^2 + (5.1)^2}$ GeV/$c \approx 13.4$ GeV/c. A sizable "missing transverse momentum" of exactly this kind, event after event, was part of the experimental signature used to identify W-boson production and decay ($W \to e^- + \bar{\nu}_e$ or $\mu^- + \bar{\nu}_\mu$) in the collider experiments of the 1980s, and remains a routine tool today for tagging any collision producing a neutrino (or, in searches for physics beyond the Standard Model, any other weakly-interacting particle that would likewise escape undetected — including, potentially, a dark-matter particle of the kind discussed in Open Questions below).

Worked Example: Threshold Energy for Pion Production

The fixed-target threshold method of Chapter 3's worked example on antiproton production applies to any particle-creation reaction, not only that one. Consider a proton beam striking a stationary proton target and producing a neutral pion, $p + p \to p + p + \pi^0$, with $m_{\pi^0}c^2 = 135.0$ MeV. Following exactly the same method, with $M = 2m_p + m_{\pi^0}$ in place of the antiproton reaction's $4m_p$:

$$(2m_p + m_{\pi^0})^2 c^4 = (E_1 + m_p c^2)^2 - (p_1 c)^2 = 2m_p^2 c^4 + 2E_1 m_p c^2. \tag{14.4}$$

Expanding the left side and solving for the beam proton's total energy E_1,

$$E_1 = m_p c^2 + 2m_{\pi^0}c^2 + \frac{(m_{\pi^0}c^2)^2}{2m_p c^2}, \tag{14.5}$$

so the threshold *kinetic* energy is

$$K_1 = E_1 - m_p c^2 = 2m_{\pi^0}c^2 + \frac{(m_{\pi^0}c^2)^2}{2m_p c^2} = 2(135.0\ \text{MeV}) + \frac{(135.0\ \text{MeV})^2}{2(938.3\ \text{MeV})} \approx 270.0\ \text{MeV} + 9.7\ \text{MeV} = 280\ \text{MeV}. \tag{14.6}$$

This is dramatically less demanding than the $6m_p c^2 \approx 5.6$ GeV threshold found for antiproton production in Chapter 3, consistent with pion physics having been accessible to the very first generation of proton accelerators in the late 1940s and early 1950s, well before the antiproton search motivated the Bevatron. Reaching successively higher thresholds like these — hundreds of MeV for pions, several GeV for antiprotons and strange particles, hundreds of GeV for the Higgs boson, and TeV-scale collision energies for direct searches beyond the Standard Model — is the entire history of why particle accelerators have grown, generation after generation, from tabletop devices to the 27-kilometer LHC.

Leptons and Quarks

The Standard Model's fundamental matter fermions fall into two families, **leptons** and **quarks**, each organized into three repeating **generations** of increasing mass, with the same gauge-interaction charges in corresponding generations. Their masses differ, and neutrino mixing means that lepton flavor is not an exact conserved label during propagation:

> The names are Greek in origin and describe relative mass, not any structural property: **lepton** comes from *leptos* ("small, thin"), fitting how light the electron and its relatives are, while **hadron** (introduced below) comes from *hadros* ("stout, bulky"), coined only in 1962 once cosmic-ray and accelerator experiments had revealed a menagerie of strongly-interacting particles considerably heavier than any lepton.

Leptons are fermions that do not feel the strong interaction. Each generation contains a charged lepton and a corresponding electrically neutral neutrino: the electron e^- and electron neutrino ν_e (generation 1, the only stable charged lepton and the only leptons found in ordinary matter); the muon μ^- and muon neutrino ν_μ (generation 2); and the tau τ^- and tau neutrino ν_τ (generation 3). The charged leptons have the same gauge-interaction charges but differ in mass and decay channels. The muon, for instance, decays via the weak interaction ($\mu^- \to e^- + \bar{\nu}_e + \nu_\mu$) with a mean lifetime of about 2.2 μs, far too short-lived to be found as a stable constituent of ordinary matter. Neutrinos have very small, nonzero masses, even though their masses are often negligible in introductory calculations.

Worked Example: Muon Decay Length in a Particle Beam

A beam of muons is produced at a particle accelerator with kinetic energy $K = 300$ MeV, well above the muon's rest energy $m_\mu c^2 = 105.7$ MeV, so the relativistic machinery of Chapter 3 is required rather than a Newtonian estimate. The total energy is $E = K + m_\mu c^2 = 405.7$ MeV, so

$$\gamma = \frac{E}{m_\mu c^2} = \frac{405.7\ \text{MeV}}{105.7\ \text{MeV}} = 3.84, \qquad \beta = \sqrt{1 - \frac{1}{\gamma^2}} = 0.965, \tag{14.7}$$

giving a lab-frame speed $v = 0.965c$. Time dilation (Chapter 2) stretches the muon's proper mean lifetime $\tau_0 = 2.20\ \mu\text{s}$ to a lab-frame mean lifetime $\tau = \gamma\tau_0 = (3.84)(2.20\ \mu\text{s}) = 8.44\ \mu\text{s}$, so the mean distance the beam travels before decaying away is

$$L = v\tau = (0.965)(3.00 \times 10^8\ \text{m/s})(8.44 \times 10^{-6}\ \text{s}) \approx 2.4 \times 10^3\ \text{m} = 2.4\ \text{km}. \tag{14.8}$$

Even with time dilation stretching the muon's lifetime nearly fourfold, one mean decay length of 2.4 km reduces the beam to $e^{-1} \approx 37\%$ of its initial intensity — a distance that accelerator and neutrino-beam facilities (which create intense neutrino beams by steering a pion beam down a long "decay pipe" and collecting neutrinos from the resulting pion decays) must design around explicitly. This is exactly the same time-dilation effect responsible for cosmic-ray muons reaching Earth's surface, applied here to a laboratory beam instead of the atmosphere.

Quarks are fermions that do feel the strong interaction, and — unlike leptons — are never observed as free, isolated particles (a phenomenon called **confinement**, discussed further below). The three generations are: up (u) and down (d) (generation 1, the constituents of ordinary protons and neutrons); charm (c) and strange (s) (generation 2); and top (t) and bottom (b) (generation 3). Quarks carry fractional electric charge, $+\frac{2}{3}e$ (u, c, t) or $-\frac{1}{3}e$ (d, s, b), the only known particles to do so, and additionally carry a strong-interaction charge called **color** (in three varieties, whimsically named red, green, and blue, with no relation to visible color), which plays a role for the strong force directly analogous to the role electric charge plays for the electromagnetic force.

For every lepton and quark there exists a corresponding **antiparticle**, with identical mass and spin but opposite electric charge (and, for quarks, opposite color) — the electron's antiparticle, the positron, was introduced already in Chapter 6's discussion of pair production, and the same particle-antiparticle structure is universal across all Standard Model fermions.

Figure 14.4 collects everything cataloged so far — three generations apiece of quarks and leptons, plus the gauge bosons of Table 14.1 and the Higgs boson introduced later in this chapter — into a single chart of the Standard Model's complete particle content.

14.3 Hadrons, Conservation Laws, and Open Questions

Hadrons and Confinement

Quarks are never seen individually; they are always found bound into composite particles called **hadrons**, a consequence of quark confinement: unlike the Coulomb or nuclear forces encountered so far, which weaken with distance, the strong force between two quarks does *not* weaken as they are pulled apart — instead, the energy stored in the strong-force field between them grows without bound, so that attempting to separate two quarks (e.g., in a high-energy collision) eventually supplies enough energy, via mass–energy equivalence, to spontaneously create a new quark-antiquark pair from the field energy itself, snapping the field into two shorter, separately confined pieces rather than yielding a single free quark. The simplest and historically most familiar hadrons occur in two configurations, both of which have net integer electric charge and (color-)neutral total color charge, consistent with confinement always producing color-neutral bound states:

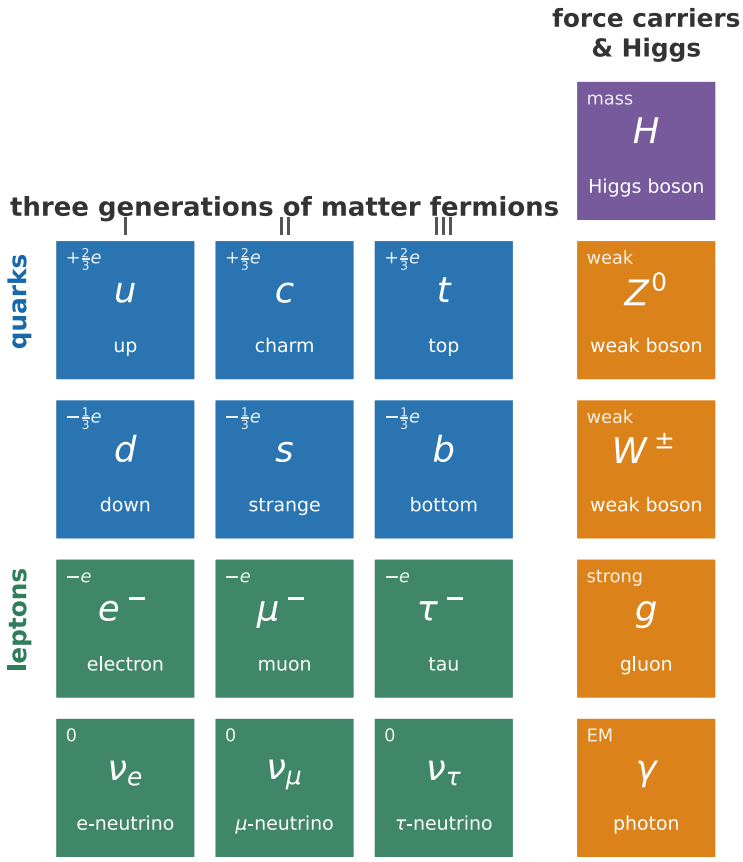

Figure 14.4: The Standard Model's fundamental particles: three repeating generations of matter fermions (quarks and leptons), plus the force-carrying gauge bosons and the Higgs boson, neither of which come in generations. Original schematic by the author, redrawn independently of the familiar CERN/Wikipedia layout.

- **Baryons**: three quarks bound together (or three antiquarks, for antibaryons). The proton (uud) and neutron (udd) are the lightest, most familiar baryons; baryons are fermions (three half-integer spins combine to a net half-integer spin) and obey a conservation law, **baryon number** (discussed below).
- **Mesons**: a quark-antiquark pair. Mesons, such as the pion ($\pi^+ = u\bar{d}$, among others), are bosons (a half-integer spin combined with a half-integer spin gives an integer net spin) and carry baryon number zero.

Quantum chromodynamics also permits color-neutral **exotic hadrons**, such as tetraquark states (two quarks and two antiquarks) and pentaquark states (four quarks and one antiquark), several of which have been observed. They are not needed for the ordinary baryon/meson classification used here, but they show that these are not the only possible arrangements of confined quarks.

The Confinement Potential: A Toy Model

The qualitative story above — that the strong-force field between two quarks stores energy without bound as they separate — can be made quantitative with a simple model that captures the essential physics. Unlike the Coulomb potential $V(r) \propto -1/r$, which weakens with distance because electric field lines spread out to fill three-dimensional space, the gluon field between a quark and an antiquark is observed, both experimentally and in lattice-QCD calculations, to collapse into a narrow, roughly constant-cross-section **flux tube**, so that separating the two ends by a distance r requires an energy that grows *linearly*,

$$V(r) \approx \sigma r, \tag{14.9}$$

where σ, the **string tension**, is measured to be about 1 GeV/fm. Once the stored energy σr climbs to a few hundred MeV — comparable to the energy needed to create a light quark-antiquark pair from the vacuum via $E = mc^2$ — it becomes energetically cheaper for the field to supply that pair than to keep stretching, so the flux tube snaps into two shorter, separately confined pieces rather than yielding an isolated quark. This typically happens after the tube has stretched only about 1 fm, comparable to a hadron's own size, which is exactly why no accelerator, however powerful, can pull a free quark out of a hadron: pushing r larger simply manufactures more hadrons, never a lone one.

Worked Example: Quark Content of the K^+ Meson

The K^+ meson (a **kaon**), like the pion above, is a quark-antiquark pair, but built from a different combination: $K^+ = u\bar{s}$, an up quark bound to a strange *antiquark*. Every antiquark carries the exact opposite electric charge of its corresponding quark, so $\bar{s}$ (the antiparticle of the $-\frac{1}{3}e$ strange quark) carries $+\frac{1}{3}e$, and

$$Q(K^+) = Q(u) + Q(\bar{s}) = +\frac{2}{3}e + \frac{1}{3}e = +e, \tag{14.10}$$

matching the kaon's experimentally known charge of $+e$ exactly. Like the pion, the K^+ is a boson with baryon number zero — but unlike the pion, it carries a nonzero **strangeness**: the strange antiquark $\bar{s}$ carries strangeness $+1$ (opposite the strange quark's $S = -1$, by the sign convention introduced below), while the up quark carries $S = 0$, giving $S(K^+) = +1$. This one extra bookkeeping number, absent from the pion, is exactly what made kaons behave so differently from pions when both were first encountered in cosmic-ray data — the subject of the next section.

Historical Context: The Particle Zoo

The tidy classification of hadrons into baryons and mesons, each built from a handful of quark flavors, is a retrospective simplification of what was, for roughly three decades, one of the most disorienting periods in the history of physics. Cosmic rays — high-energy particles arriving from space and colliding with nuclei in the upper atmosphere — were, before particle accelerators reached comparable energies, the only available source of exotic new particles, and they did not disappoint. In 1936–1937, Carl Anderson and Seth Neddermeyer, studying cosmic-ray tracks in a cloud chamber, identified a new particle with a mass about 200 times the electron's — initially mistaken for the meson Hideki Yukawa had predicted the previous year as the carrier of the nuclear force, but eventually recognized (it interacted far too weakly, and penetrated far too much matter, to be Yukawa's particle) as an entirely new, heavier cousin of the electron: the **muon**, already met above. Yukawa's actual meson turned up a decade later, in 1947, when Cecil Powell, Giuseppe Occhialini, and César Lattes, examining photographic emulsions exposed at high-altitude observatories, identified the **pion**, caught in the act of decaying into exactly the muon Anderson and Neddermeyer had already found — untangling, at a stroke, two distinct particles that had been conflated for a decade. That same year, George Rochester and Clifford Butler, also working from cosmic-ray cloud-chamber photographs, found still stranger tracks: V-shaped pairs of particle trails, produced abundantly yet decaying only slowly — the signature of what came to be called, aptly, **strange particles**: the kaons of the worked example above, and the lightest hyperons (baryons heavier than the proton and neutron, such as the Λ^0 introduced earlier in this chapter).

By the 1950s, particle accelerators had overtaken cosmic rays as the primary discovery tool, and the pace of discovery, if anything, accelerated: dozens of "elementary" hadrons were cataloged through that decade and into the next, with no organizing principle in sight — derisively called the **particle zoo** by physicists who had expected nature's fundamental building blocks to be few in number, not a sprawling menagerie. Order emerged in 1961, when Murray Gell-Mann and, independently, Yuval Ne'eman noticed that plotting hadrons by electric charge against strangeness placed each family into strikingly regular geometric patterns — hexagons and triangles of eight or ten particles apiece — a scheme Gell-Mann named the **eightfold way**. The scheme's predictive power was demonstrated dramatically in 1964, when a particle it predicted but which had not yet been observed, the Ω^- baryon, was found at Brookhaven National Laboratory with essentially the predicted mass, charge, and strangeness. That same year, Gell-Mann and, independently, George Zweig proposed the deeper explanation for why the pattern worked at all: every hadron in the eightfold way's patterns is simply a bound state of a small number of still more fundamental constituents, which Gell-Mann named **quarks** — the up, down, and strange quarks already introduced in this chapter, with the heavier charm, bottom, and top quarks discovered only later, between 1974 and 1995, as accelerators reached the correspondingly higher energies needed to produce them. What had looked, in the 1950s, like a hopeless proliferation of unrelated "elementary" particles turned out to be nothing more than the many possible ways of combining a mere handful of truly fundamental building blocks — precisely the kind of underlying simplicity the Standard Model was built to express.

Predicting the Unknown from a Pattern

The eightfold way's success in predicting the Ω^- before it was ever observed is not a one-off in the history of physics. Chapter 11 recounts how Moseley's law, another purely

empirical pattern relating a measured quantity ($\sqrt{f}$, there; charge and strangeness, here) to an integer index, correctly predicted the existence of three then-undiscovered elements ($Z = 43$, 61, 75) from gaps in an otherwise orderly sequence. In both cases, an organizing pattern was trusted well before anyone understood *why* it held — the "why" (electron shell structure in one case, quark substructure in the other) came only afterward.

Where the Name "Quark" Came From

Gell-Mann's name for the new constituents was borrowed, by his own account, from a nonsense line in James Joyce's *Finnegans Wake* — "Three quarks for Muster Mark!" — which happened to fit neatly with the three quark flavors (u, d, s) known at the time. Zweig, who proposed the same constituents independently and simultaneously, called them "aces" instead; Gell-Mann's name, and his convention of fractional-e electric charges, is the one the entire field still uses today.

Conservation Laws

Not every combination of particles satisfying energy-momentum conservation is actually observed to occur; particle reactions additionally obey several conservation laws, some familiar from earlier chapters and some new to particle physics:

- **Electric charge** is conserved in every known interaction, without exception.
- **Baryon number** B (defined as +1 for each baryon, -1 for each antibaryon, 0 for all other particles, including mesons and leptons) is conserved in every observed reaction — this is why, for instance, the proton, the lightest baryon, is observed to be stable (or at least extremely long-lived: no proton decay has ever been observed, despite dedicated searches, placing its lifetime, if it decays at all, above 10^{34} years), since there is no lighter baryon for it to decay into consistent with B conservation.
- **Lepton number**, separately for each generation (L_e, L_μ, L_τ, each +1 for the corresponding particle, -1 for its antiparticle, 0 otherwise), is conserved at the individual weak-interaction vertices used in this chapter and is an excellent approximation in most particle decays. Neutrino oscillations show that these flavor-specific numbers are not exact over long propagation distances. This is why, for example, muon decay $\mu^- \to e^- + \bar{\nu}_e + \nu_\mu$ produces *both* an electron antineutrino and a muon neutrino, rather than either alone, to separately conserve L_e and L_μ at the decay vertices: the initial state has $L_\mu = +1, L_e = 0$, and the combination $\bar{\nu}_e + \nu_\mu$ on the right-hand side reproduces $L_e = 0$, $L_\mu = +1$. A flavor-changing decay such as $\mu^- \to e^- + \gamma$ is therefore forbidden at the ordinary tree level and is extraordinarily suppressed once neutrino masses and mixing are included.

These conservation laws function exactly as energy, momentum, and angular momentum conservation do in earlier chapters: a proposed reaction consistent with all other physics can nonetheless be immediately ruled out if it violates one of these rules, and they provide a fast, purely bookkeeping-based check on whether an observed or hypothesized particle process is allowed.

Strangeness: An (Almost) Conserved Quantum Number

The particle-zoo puzzle described above is also where **strangeness** S enters as a genuine, bookkeepable quantum number, alongside charge, baryon number, and lepton number. Strangeness is assigned by quark content: $S = -1$ for each strange quark s a particle contains, $S = +1$ for each strange antiquark $\bar{s}$, and $S = 0$ for every other quark flavor (a sign convention fixed historically, before quarks themselves were proposed, when strangeness was assigned directly to particles from their production and decay patterns rather than derived from a quark that had not yet been identified). The defining, and at first deeply puzzling, experimental fact about strangeness is that it is treated differently by different interactions:

> **Warning**
> The sign of strangeness is historical, not intuitive: a strange quark carries $S = -1$, while a strange antiquark carries $S = +1$. "Almost conserved" means exactly conserved by the strong and electromagnetic interactions, but not by the weak interaction.

- The **strong** and **electromagnetic** interactions conserve strangeness exactly, in every observed reaction.
- The **weak** interaction does not: a single tree-level charged-current vertex can change strangeness by one unit, consistent with a $W^{\pm}$ vertex converting an s quark to an up-type quark, or vice versa (the same charge-budget argument given above for why only the weak interaction changes quark flavor at all). Higher-order weak processes can produce larger net changes, but they are strongly suppressed.

This single fact resolves Rochester and Butler's original puzzle: strange particles were produced copiously but decayed reluctantly because **associated production**, the strong-interaction process that actually creates strange particles in accelerator and cosmic-ray collisions (worked through below), always creates strangeness in matched pairs summing to the initial strangeness of zero — consistent with strong-interaction conservation, and proceeding on the characteristically fast strong-interaction timescale of order 10^{-23} s. But once created, an isolated strange particle such as the Λ^0 has no strangeness-*conserving* decay available to it: there is no lighter, non-strange combination of ordinary particles it can fall apart into without changing S. It can only decay via the strangeness-violating weak interaction, on the correspondingly slow weak-interaction timescale — typically 10^{-10} s, some thirteen orders of magnitude longer than a typical strong-interaction process, though still far too brief to see with the naked eye. Produced fast, decaying slow: hence "strange."

> **A Fast Checklist for Particle Reactions**
> When checking whether a proposed reaction is allowed, work through the conservation laws in order of how absolute they are, not in the order this chapter introduced them. Charge and baryon number are conserved without exception in every observed process, so check those first — a violation there kills the reaction outright, with no further work needed. Lepton number, checked separately per generation, is next for the elementary vertices considered here, with the caveat that neutrino oscillations can change the flavor labels during propagation. Only then check strangeness, and check it last precisely because it is *allowed* to change: a reaction with $\Delta S = 0$ can proceed strongly or electromagnetically, while $\Delta S = \pm 1$ at tree level signals a weak process. Larger changes are possible only

through higher-order weak processes and are strongly suppressed, not absolutely forbidden. The size of ΔS, in other words, is a useful fast diagnostic for which interaction a reaction must proceed through.

Worked Example: Strangeness Conservation in Associated Production and Decay

Consider two reactions involving the Λ^0 (uds) introduced earlier in this chapter and the K^0 meson ($d\bar{s}$):

$$\pi^- + p \to K^0 + \Lambda^0 \qquad \text{(associated production)}, \qquad \Lambda^0 \to p + \pi^- \qquad \text{(the } \Lambda^0\text{'s actual decay)}. \tag{14.11}$$

Neither the pion nor the proton on the left of the first reaction contains a strange quark ($\pi^- = \bar{u}d$, $p = uud$), so the total initial strangeness is $S_i = 0$. On the right, $K^0 = d\bar{s}$ carries $S = +1$ (from its $\bar{s}$) and $\Lambda^0 = uds$ carries $S = -1$ (from its s), for a total final strangeness $S_f = (+1)+(-1) = 0$. Strangeness is conserved — consistent with this reaction proceeding via the strong interaction, exactly the associated-production mechanism described above, with the K^0 and Λ^0 created together, their strangeness values canceling. (Charge and baryon number are conserved as well: Q: $-e + e = 0 = 0 + 0$; B: $1 + 0 = 0 + 1$.)

The second reaction, the Λ^0's actual decay, is different: $S_i = S(\Lambda^0) = -1$, while $S_f = S(p) + S(\pi^-) = 0 + 0 = 0$, a change of $\Delta S = +1$ — forbidden for the strong or electromagnetic interaction, but allowed, one unit at a time, for the weak interaction. This is consistent with the Λ^0's measured mean lifetime of about 2.6×10^{-10} s, squarely in the weak-interaction range, roughly 10^{13} times longer than the strong-interaction timescale on which the very same particle was created in the first reaction.

The Higgs Mechanism

A long-standing puzzle in the Standard Model was that the mathematical framework describing the weak and electromagnetic interactions in a unified way most naturally predicts that all fundamental particles, including the electron and quarks, should be massless — directly contradicted by experiment. The resolution, proposed independently by several theorists in 1964 (and associated most closely with the name of Peter Higgs), is that space is permeated by a nonzero background field, the **Higgs field**, and that fundamental particles acquire mass through their interaction with this field: a particle that couples strongly to the Higgs field behaves as though it has large inertia (large mass) as it moves through the field, while a particle that does not couple to it (the photon, for instance) remains exactly massless. Associated with the Higgs field, exactly as the electromagnetic field has an associated particle (the photon) representing its quantized excitations, is the **Higgs boson**, whose discovery at the Large Hadron Collider in 2012 — decades after it was first predicted, and requiring a purpose-built accelerator capable of reaching the multi-hundred-GeV collision energies needed to produce it directly — provided direct experimental confirmation of the mechanism and completed the last missing piece of the Standard Model's particle content.

The strength of a given particle's coupling to the Higgs field is not itself predicted by the mechanism — it is a separate, measured number for every fermion and for the $W^{\pm}/Z^0$ bosons — but the *pattern* it produces is a striking, directly testable success: the more strongly a particle couples, the larger its mass, and the Standard Model's heaviest fermion, the top

quark ($m_t c^2 \approx 173$ GeV, almost the mass of an entire gold nucleus packed into one point particle), has the largest known Yukawa coupling, of order one. The electron, some 300,000 times lighter, couples correspondingly more weakly. The photon's exact masslessness is, in this same language, simply the statement that the photon has no mass-generating coupling to the Higgs field — a consequence of the specific way electroweak symmetry breaking mixes the underlying gauge bosons of the unified electroweak theory into the photon and the $W^{\pm}/Z^0$ observed today, with only the latter three acquiring mass from the mechanism.

The Higgs Field Explains Mass — But Not Most of Your Mass

It is tempting to read the Higgs mechanism as a complete explanation of why matter has mass, but for ordinary matter it accounts for only a small fraction of it. The Higgs mechanism generates the *fundamental* masses cataloged in Figure 14.4 — the rest mass of an individual quark or electron, on its own, with nothing else around it. A proton's mass is not simply the sum of its three quarks' Higgs-generated masses: two up quarks and a down quark contribute barely 1% of the proton's 938.3 MeV. The remaining 99% is energy in the quark–gluon field, including confinement and the quarks' relativistic motion, converted to rest mass via $E = mc^2$ exactly as in the nuclear mass defect of Chapter 13. The Higgs field is why quarks and electrons have mass at all; quantum chromodynamics, not the Higgs field, accounts for almost all of the mass of everyday matter.

Open Questions

As the final chapter of this book, it is worth being explicit that the Standard Model, for all its precision and predictive success — every particle in its minimal particle content has now been found, and many of its predictions for their properties agree with experiment to several decimal places — is not a complete "theory of everything." Several major open questions define the current frontier of particle physics and cosmology, worth naming even though, unlike everything else in this book, none has yet been resolved:

- **Dark matter.** Galaxies and clusters of galaxies rotate and move as though they contain roughly five times more gravitating mass than is visible in stars, gas, and dust combined — a discrepancy first noted from galaxy rotation curves and since confirmed by multiple independent lines of evidence (gravitational lensing, the large-scale structure of the universe, the cosmic microwave background). No known Standard Model particle has the right properties — electrically neutral, stable or extremely long-lived, interacting only very weakly with ordinary matter — to account for it, and despite decades of dedicated direct-detection experiments, no dark-matter particle has yet been observed in the laboratory.
- **Matter–antimatter asymmetry.** Chapter 6 and this chapter both treat particles and antiparticles as, in most respects, perfectly symmetric, yet the observable universe is made overwhelmingly of matter, with no evidence anywhere of large antimatter regions — even though the Big Bang, run forward through known physics, should have produced matter and antimatter in almost exactly equal amounts, which would then have mutually annihilated into pure radiation and left nothing behind to form galaxies, stars, or the reader of this sentence. Explaining the tiny observed excess of matter over antimatter (**baryogenesis**) requires physics that violates matter–antimatter symmetry more thoroughly than the Standard Model's confirmed interactions currently account for.



Figure 14.5: The same point as the warning box above, developed further: most of the mass in the room around you is not Higgs mass at all. Watch on YouTube: Your Mass is NOT From the Higgs Boson

- **Neutrino mass.** The Standard Model outlined in this chapter, in its original form, predicted neutrinos to be exactly massless, like the photon. Neutrino oscillation experiments — a neutrino produced in one flavor (say ν_μ) measured, some distance later, with a nonzero probability of being detected as a different flavor (ν_e or ν_τ) — demonstrated conclusively, beginning in the late 1990s, that neutrinos have a small but nonzero mass: a firmly established experimental fact with no settled place in the version of the Standard Model presented in this chapter, and one of the clearest concrete signs that the model, though successful, is not the final word.
- **Unification.** The electromagnetic and weak interactions are already understood as two facets of a single underlying "electroweak" interaction, unified at sufficiently high energy — a success this chapter's Higgs mechanism is part of. Many physicists suspect the strong interaction unifies with the electroweak interaction at a still higher energy (**grand unification**), and that gravity itself — entirely absent from the Standard Model, and still described only by Einstein's classical general relativity — must eventually be brought into a single quantum-mechanical framework together with the other three forces. Such a "theory of everything" remains, despite decades of effort (string theory and loop quantum gravity among the leading proposals), unconfirmed by any experiment.

These are not signs of failure so much as an honest map of where the frontier currently sits: the Standard Model surveyed in this chapter is, by a wide margin, the most thoroughly tested theory in the history of physics, and every one of these open questions is being actively pursued — at facilities from the LHC to underground dark-matter detectors to neutrino observatories — by the direct descendants of the same experimental methods, particle accelerators, colliders, and increasingly sophisticated layered detectors, introduced earlier in this chapter.

If you want the shortest reading of this chapter's ending, it is this: the Standard Model is extraordinarily precise, but it still does not explain dark matter, the matter-antimatter asymmetry, neutrino masses, or how gravity joins the quantum forces.

14.4 Summary

- Particles are classified by spin as **fermions** (half-integer spin, obey the exclusion principle; matter is built from these) or **bosons** (integer spin, no exclusion restriction; forces are mediated by these).
- Four fundamental interactions — **strong**, **electromagnetic**, **weak**, and **gravitational** — are distinguished by relative strength, range, and mediating gauge boson (gluon, photon, $W^\pm/Z^0$, and the hypothesized graviton, respectively).
- Matter fermions are **leptons** (no strong interaction; electron, muon, tau and their neutrinos, across three generations) and **quarks** (feel the strong interaction, carry fractional charge and color, never observed free due to **confinement**).
- Quarks bind into color-neutral **hadrons**, most familiarly **baryons** (three quarks, e.g. proton uud, neutron udd) and **mesons** (quark-antiquark pairs); exotic tetraquark and pentaquark states are also observed.
- **Conservation laws** — electric charge, baryon number, and (at elementary vertices, to an excellent approximation) lepton number — determine which proposed particle reactions are physically allowed. **Strangeness** is a further, only-approximately-conserved quantity: exactly conserved by the strong and electromagnetic interactions, but changeable by one unit at a tree-level charged-current weak vertex, which is why strange particles are produced quickly (via the strong interaction) but decay slowly (via the weak).

Higher-order weak processes can change it by more than one unit, but are strongly suppressed.

- The **Higgs field**, and its associated **Higgs boson** (discovered 2012), is the mechanism by which most fundamental Standard Model particles acquire mass.
- A **Feynman diagram** depicts a particle interaction as fermion lines (particle arrows forward in time, antiparticle arrows reversed) meeting at **vertices**, connected by internal, virtual boson lines; every exactly-conserved quantity must balance at each individual vertex, and only a charged-current $W^{\pm}$ vertex directly changes a quark's flavor, since it alone carries enough charge to balance an up-type/down-type quark transition.
- **Particle accelerators** (linear or, more commonly, synchrotron-based) collide beams either against a fixed target or head-on in a **collider**, and layered **detectors** — tracking, calorimetry, and muon identification — reconstruct what each collision produced; the Large Hadron Collider at CERN is the largest such machine yet built.
- The **particle zoo** of hadrons discovered via cosmic rays and early accelerators (muon, 1936; pion and strange particles, 1947) was organized by the **eightfold way** (Gell-Mann and Ne'eman, 1961) and explained by the **quark model** (Gell-Mann and Zweig, 1964).
- Gravity is negligible for individual particles: the ratio of the gravitational to the electromagnetic force between two protons is only $\sim 10^{-36}$, because mass (unlike charge) is always attractive and accumulates without limit, while ordinary bulk matter is nearly charge-neutral.
- The Standard Model remains incomplete: **dark matter**, the **matter–antimatter asymmetry**, nonzero **neutrino mass**, and the search for a **unified** theory including gravity are among the major open questions at the current frontier of particle physics.

14.5 Check Your Understanding

Five short, auto-graded questions cycle within one compact activity. The set mixes multiple choice, true/false, drag-the-words, fill-in-the-blank, and mark-the-words formats, and opens on the Standard Model chart.

1. **Multiple choice, from a figure.** Which Standard Model interaction is carried by gluons? Figure description: Chart of Standard Model matter particles and force-carrying bosons. Choices: (A) The strong interaction. (B) The electromagnetic interaction. (C) The weak interaction. (D) Gravity.

2. **True or false.** A meson is made from three quarks, while a baryon is a quark–antiquark pair.

3. **Drag the words.** Complete the Standard Model classifications. Matter particles with half-integer spin are _____. Force carriers are _____. Three-quark hadrons are _____. Quark–antiquark hadrons are _____. (Terms: fermions; bosons; baryons; mesons; leptons only; gauge fermions.)

4. **Fill in the blanks.** Complete the proton and detector statements. A proton has quark content _____, and missing transverse momentum can signal an unseen _____.

5. **Mark the words.** Mark the two quantities that must balance in every allowed particle reaction. "Every reaction must conserve electric charge and baryon number. Lepton number is also conserved to excellent approximation in Standard Model reactions."

14.6 Conceptual Questions

1. Distinguish elementary particles from composite hadrons using quarks and leptons as examples.

2. What does an internal line in a Feynman diagram represent, and why should it not be read as a literal photographed trajectory?

3. Why do colliders devote separate detector layers to tracking, calorimetry, and muon identification?

4. Explain how fractional quark charges combine to give integer charges for observed baryons and mesons.

5. Why does confinement prevent an isolated quark from being collected even in a very energetic collision?

6. What role does the Higgs field play in the Standard Model, and what kinds of mass does it not by itself explain?

14.7 Problems

Exercise 14.1
(Straightforward)

Classify each of the following as a fermion or boson, based on its spin: photon ($s = 1$), electron ($s = \frac{1}{2}$), pion ($s = 0$), proton ($s = \frac{1}{2}$).

Exercise 14.2
(Straightforward)

Determine the electric charge of a baryon composed of uds (this particle is the Λ^0) using the quark charges $+\frac{2}{3}e$ for u, $-\frac{1}{3}e$ for d and s, and check your result against the known charge of the Λ^0 (zero).

Exercise 14.3
(Straightforward)

Determine whether each proposed reaction conserves charge, baryon number, and lepton number as required, and state which conservation law (if any) forbids the ones that are not allowed: (a) $p \to e^+ + \gamma$, (b) $n \to p + e^- + \bar{\nu}_e$, (c) $\mu^- \to e^- + \gamma$, (d) $p + p \to p + p + \pi^0$.

Exercise 14.4
(Straightforward)

Using the known quark content of the proton (uud) and neutron (udd), and the quark charges given in Exercise 14.2, verify that the proton has charge $+e$ and the neutron has charge 0.

Exercise 14.5
(Moderate)

The muon decays via $\mu^- \to e^- + \bar{\nu}_e + \nu_\mu$. Explain, using lepton-number conservation applied separately to the electron-generation number L_e and muon-generation number L_μ, why the decay $\mu^- \to e^- + \gamma$ alone (without the two neutrinos) is forbidden at the ordinary tree level, even though it conserves charge, energy, and momentum.

Exercise 14.6
(Moderate)

Explain, in your own words, why quark confinement means that the constituent quarks of a proton can never be observed as free, isolated particles no matter how much energy is used to try to separate them, and contrast this with the behavior of the electromagnetic force between two separated electric charges, which weakens (rather than growing) with increasing separation.

Exercise 14.7
(Moderate)

The K^- meson has quark content $K^- = \bar{u}s$. (a) Using the quark charges given in Exercise 14.2 and the rule that an antiquark carries the opposite charge of its quark, find the electric charge of K^-. (b) Find its strangeness. (c) Explain why your results show that K^- is the antiparticle of the K^+ found in the worked example on kaon quark content, and state the general rule (in terms of charge and strangeness) that a particle and its antiparticle must always satisfy.

Exercise 14.8
(Moderate)

Determine whether the reaction $K^- + p \to \Lambda^0 + \pi^0$ conserves charge, baryon number, and strangeness, and state whether it can proceed via the strong interaction, the weak interaction, or neither.

Exercise 14.9
(Challenging)

The muon decays via $\mu^- \to e^- + \bar{\nu}_e + \nu_\mu$ (Chapter 2), and, like neutron beta decay, this proceeds by μ^- emitting a virtual W^- and converting directly into ν_μ at one vertex, with the W^- then decaying to $e^- + \bar{\nu}_e$ at a second vertex. Sketch this two-vertex diagram in the style of Figure 14.2, and verify explicitly that electric charge and each generation's lepton number (L_e and L_μ separately) balance at *each* vertex individually, following the method of the worked example on reading a Feynman diagram.

Exercise 14.10
(Challenging)

A second muon beam is produced with total energy $E = 1.20$ GeV (using $m_\mu c^2 = 105.7$ MeV and proper mean lifetime $\tau_0 = 2.20$ μs, as in the worked example on muon decay length). Find (a) γ, (b) β, and (c) the mean decay length of this beam in the lab frame. Compare your answer to the 2.4 km found in the worked example, and explain the direction of the difference in terms of γ.

Exercise 14.11

(Challenging)

Repeat the gravitational-versus-electromagnetic force-ratio calculation of the worked example in "The Four Fundamental Interactions," but for two electrons instead of two protons (using $m_e = 9.109 \times 10^{-31}$ kg in place of m_p, with G, k, and e unchanged). Explain, in terms of the formula for the ratio, why the result is so much smaller than the two-proton ratio even though an electron and a proton carry exactly the same magnitude of electric charge.

Chapter 15

Teaching Guide

This guide is for an instructor who wants a fast, workable 10-week path through the book. It keeps the narrative spine intact: relativity, wave optics, quantum mechanics, atomic structure, then nuclear and particle physics. If you have more time, stretch any week into two; if you have less, trim from the right-hand side of the table first.

15.1 Suggested 10-Week Pacing

Week	Assign	Purpose
1	Chapter 1, all sections	Remove the ether picture and motivate special relativity.
2	Chapter 2, “Simultaneity, Time Dilation, and Length Contraction”	Establish the new kinematics.
3	Chapter 2, the remaining sections; Chapter 3, “Relativistic Momentum”	Finish the Lorentz transformation and start dynamics.
4	Chapter 3, the remaining sections	Cover energy, four-vectors, and threshold arguments.
5	Chapter 4, the first two sections	Build interference, coherence, and superposition.
6	Chapter 4, the remaining section; Chapter 5, the first section	Finish interference and begin diffraction.
7	Chapter 5, the remaining sections; Chapter 6, the first section	Finish wave optics and introduce photon behavior.
8	Chapters 6 and 7	Cover the photoelectric effect, Compton scattering, de Broglie waves, and uncertainty.
9	Chapters 8 and 9	Develop the Schrödinger equation and 3D quantum mechanics.
10	Chapters 10 and 11, then selected reading from Chapters 12-14	End with atoms, many-electron structure, and a brief look at nuclei and the Standard Model.

15.2 What To Trim First

If the term gets tight, the best places to compress are:

- Chapter 5, later sections on gratings and resolution.
- Chapter 9, the isotropic oscillator material.
- Chapter 12, the more detailed molecular-orbital discussion.
- Chapter 14, the historical particle-zoo section.

If you want a course that reaches the Standard Model in less than a full semester, keep Chapters 1-11 in order, then assign only the opening section of Chapter 13 and the first and last sections of Chapter 14.

Chapter 16

Glossary

Terms are defined here as they are used in this book. Abbreviations that are only expansions are handled by hover text in the chapters and are not repeated.

Chapter 17

Bibliography

This bibliography collects the principal texts, historical papers, and data references used for further reading and fact-checking. Content licensing and adaptation provenance are recorded separately in Sources & Attribution.

Broad treatments used across the book include OpenStax and standard modern, quantum, and nuclear-physics texts [OpenStax, 2016, Beiser, 2003, Eisberg and Resnick, 1985, Griffiths and Schroeter, 2018, Krane, 1987]. The historical foundations are represented by the Michelson–Morley and Einstein relativity papers [Michelson and Morley, 1887, Einstein, 1905b], the early quantum papers of Planck, Einstein, Compton, Schrödinger, Born, and Heisenberg [Planck, 1901, Einstein, 1905a, Compton, 1923, Schrödinger, 1926, Born, 1926, Heisenberg, 1927], and decisive experiments on angular momentum and matter-wave interference [Gerlach and Stern, 1922, Davisson and Germer, 1927, Colella et al., 1975, Arndt et al., 1999]. Current atomic, nuclear, and particle data are anchored by the NIST Atomic Spectra Database, the Atomic Mass Evaluation, and the Particle Data Group review [Kramida et al., 2024, Wang et al., 2021, Navas and others, 2024].

Acronyms

CERN European Organization for Nuclear Research. 29, 302, 303, 307, 316

CM center of momentum. 51–53, 56, 59, 302

COW Colella-Overhauser-Werner. 148, 168

GPS Global Positioning System. 35, 38

LCAO linear combination of atomic orbitals. 251, 254

LIGO Laser Interferometer Gravitational-Wave Observatory. 88, 93

MO molecular orbital. 251–253, 257, 266

PET positron-emission tomography. 51, 130, 135, 141

STM scanning tunneling microscope. 181, 187

UV ultraviolet. 139, 264

VSEPR valence-shell electron-pair repulsion. 249, 250

Bibliography

M. Arndt, O. Nairz, J. Vos-Andreae, C. Keller, G. van der Zouw, and A. Zeilinger. Wave-Particle Duality of C60 Molecules. *Nature*, 401:680–682, 1999. doi: 10.1038/44348.

A. Beiser. *Concepts of Modern Physics*. McGraw-Hill, 6 edition, 2003.

M. Born. Zur Quantenmechanik der Stoßvorgänge. *Zeitschrift für Physik*, 37:863–867, 1926. doi: 10.1007/BF01397477.

R. Colella, A. W. Overhauser, and S. A. Werner. Observation of Gravitationally Induced Quantum Interference. *Physical Review Letters*, 34:1472–1474, 1975. doi: 10.1103/PhysRevLett.34.1472.

A. H. Compton. A Quantum Theory of the Scattering of X-Rays by Light Elements. *Physical Review*, 21:483–502, 1923. doi: 10.1103/PhysRev.21.483.

C. J. Davisson and L. H. Germer. Diffraction of Electrons by a Crystal of Nickel. *Physical Review*, 30:705–740, 1927. doi: 10.1103/PhysRev.30.705.

A. Einstein. On a Heuristic Point of View Concerning the Production and Transformation of Light. *Annalen der Physik*, 17:132–148, 1905a. doi: 10.1002/andp.19053220607.

A. Einstein. On the Electrodynamics of Moving Bodies. *Annalen der Physik*, 17:891–921, 1905b. doi: 10.1002/andp.19053221004.

R. Eisberg and R. Resnick. *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles*. Wiley, 2 edition, 1985.

W. Gerlach and O. Stern. Der experimentelle Nachweis der Richtungsquantelung im Magnetfeld. *Zeitschrift für Physik*, 9:349–352, 1922. doi: 10.1007/BF01326983.

D. J. Griffiths and D. F. Schroeter. *Introduction to Quantum Mechanics*. Cambridge University Press, 3 edition, 2018. doi: 10.1017/9781316995433.

W. Heisenberg. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, 43:172–198, 1927. doi: 10.1007/BF01397280.

A. Kramida, Y. Ralchenko, J. Reader, and NIST ASD Team. Nist Atomic Spectra Database, 2024. URL https://physics.nist.gov/asd.

K. S. Krane. *Introductory Nuclear Physics*. Wiley, 1987.

A. A. Michelson and E. W. Morley. On the Relative Motion of the Earth and the Luminiferous Ether. *American Journal of Science*, 34:333–345, 1887. doi: 10.2475/ajs.s3-34.203.333.

S. Navas and others. Review of Particle Physics. *Physical Review D*, 110:030001, 2024. doi: 10.1103/PhysRevD.110.030001.

OpenStax. *University Physics, Volume 3*. OpenStax, Rice University, Houston, 2016. URL https://openstax.org/details/books/university-physics-volume-3.

M. Planck. On the Law of Distribution of Energy in the Normal Spectrum. *Annalen der Physik*, 4:553–563, 1901. doi: 10.1002/andp.19013090310.

E. Schrödinger. Quantisierung als Eigenwertproblem. *Annalen der Physik*, 79:361–376, 1926. doi: 10.1002/andp.19263840404.

M. Wang, W. J. Huang, F. G. Kondev, G. Audi, and S. Naimi. The AME 2020 Atomic Mass Evaluation, 2021. Chinese Physics C 45, 030003.