PH 447 Lecture Notes

E. Tilton

Fall 2020

Acknowledgements & Disclaimer

Large portions of this course are based upon a course that I took in graduate school. It was designed by N. W. Halverson, who has kindly given me permission to adapt it to an undergraduate course at Regis University. My personal experience of the material was also heavily shaped by N. Schneider, who taught the section of the course that I took as a student. Anything that works well in this course should probably be attributed to them. N. W. Halverson also drew from many sources himself, as his original acknowledgement from the course notes:

> I have drawn on many sources in compiling these lecture notes. I have shamelessly purloined material from excellent courses on related topics from my graduate career at Caltech, namely Dr. Thomas Gottschalk's *Physics Math Methods: Statistics* course, and Prof. John Carlstrom's *Astronomical Instrumentation* course. In addition, I have drawn on material from many texts, including Wall & Jenkins, *Practical Statistics for Astronomers*; Lupton, *Statistics in Theory and Practice*; Mclean, *Electronic Imaging in Astronomy: Detectors and Instrumentation*; Press et al, *Numerical Recipes in C*; and Bevington & Robinson, *Data Reduction and Error Analysis for the Physical Sciences.* I am also indebted to members of the APS department and others who provided lecture material, helped develop data analysis projects, and engaged in useful discussions: Nick Schneider, John Bally, Dick McCray, Jeff Linsky, Charles Danforth, Bob Ergun, Jeremy Darling, Larry Esposito, Web Cash, and Jay Austermann (U. Mass, now in CASA). Thanks also go to Adam Ginsburg and Devin Silvia who submitted corrections and suggestions for the 2008 version of these notes.
>
> These notes are still a work in progress. Purloined material is not properly acknowledged. There are certainly errors; please bring particularly egregious ones to my attention. I do hope these notes will serve as a reference for a course in which no single textbook sufficiently covers the broad range of advanced material.

Contents

Acknowledgements & Disclaimer i

1 Course Overview 1

1.1 Logistical Preliminaries . . . 1

1.2 Course Overview . . . 1

2 Mathematical Preliminaries 3

2.1 Basic Vector Operations . . . 3

2.1.1 Vectors . . . 3

2.1.2 Vector Operations . . . 4

2.2 Derivatives and the Del Operator . . . 6

2.2.1 Ordinary and Partial Derivatives . . . 6

2.2.2 The Del Operator . . . 7

2.2.3 Anything else we need? . . . 9

3 Waves 10

3.1 The Wave Equation . . . 10

3.2 Sinusoidal Waves . . . 13

3.2.1 Defining the Sinusoidal Wave . . . 13

3.2.2 Waves in Complex Notation . . . 15

3.2.3 Linear Combinations of Sinusoidal Waves . . . 16

4 Electromagnetic Waves 17

4.1 The Wave Equation for **E** and **B** . . . 17

4.2 The Electrodynamic Boundary Conditions . . . 19

4.3 The Reflection and Transmission of Light 21

5 The Basic Ideas of Geometric Optics **24**

5.1 Rays and Reversibility in Geometric Optics 24

5.2 Reflected and Refracted Images in Planes 24

6 Basic Lenses and Mirrors: The Spherical Case and the Thin Lens Equation **29**

6.1 More Complicated Optical Surfaces and Their Limitations 29

6.2 Spherical Optical Elements as Discussed by Pedrotti, Pedrotti, & Pedrotti . 30

Bibliography **43**

Chapter 1

Course Overview

1.1 Logistical Preliminaries

This course will exist in a variety of forms, including written work, computer programming, and our interpersonal interactions. Each of these carries with it its own set of hurdles, skills, and rules. Please take a look at our syllabus, which is a separate document, to better understand how this course will be structured. Within that document, you will probably begin to get a feel for my personal teaching philosophy. One important aspect of that philosophy is that I expect active participation at all times: when you are reading, when we are having class, when we are working on a group project, etc. You only learn physics by doing it, so as you read these notes, please remember to do so actively, thinking through the implications of every statement and doing scratch work where necessary.

1.2 Course Overview

What is this course? It's title is "Optics, Observations, and Analysis Techniques," but it's worth asking why those three somewhat-distinct topics have been grouped together.

In many scientific experiments, or even entire scientific disciplines, answering a scientific question involves observing something. Though this observation could be metaphorical (a computer might be tallying something, for example), it is also often completely literal (we look at the light coming from something). Nearly every scientific disipline, then, requires understanding how light is affected by optical instrumentation, whether that intrument is a camera attached to a microscope or just an old-fashioned human eye. But even once an observation is made, how ought we interpret it? To do so, we must understand how to transform our observations into quantifiable data, which we we can only place in real-world context via the careful application of statistical reasoning.

My own field of research – astronomy – is perhaps the most extreme form of this observational paradigm. When I study a galaxy, I cannot poke it with a stick, or visit it, or isolate it. I can only observe it, as it appears right now, using light. Though there are a few exceptions to these restrictions (visiting planets in our own solar system, observing via neutrinos, etc.), they generally restrict all science beyond our own planet – i.e., most topics in the universe! In this course, therefore, we will take astronomy as our case study, thinking about telescopes

and distant objects. But a small telescope pointed down is just a microscope, and math works in any discipline, so do not mistake our examples as a lack of general applicability!

In the broadest sense, then, our goals for this course are to

- Understand the fundamentals of how instruments at multiple wavelengths of light collect and affect data
- Understand principles and limitations of statistical inference and data analysis techniques
- Derive physical measurements and uncertainties with hands-on analysis of real datasets.

This course is a new experiment at Regis, and it will be taught via some unusual formats, so we may not perfectly acheive all aspects of these goals. But that's OK! This course is, for the most part, highly practical, so hopefully whatever we do get through will be useful.

Let's get started. We'll begin with a discussion of light as a wave.

Chapter 2

Mathematical Preliminaries

Before we can tackle the wave-like nature of light, we need to make sure that we are all equally confident with the mathematical tools that we will need. While some of these should be familiar from the physics and calculus that you have taken so far, some will be new if you have not yet taken courses in vector calculus or differential equations. Most of these mathematical facts will not be used frequently in this class, they are presented here as a reference for the few times that we will need them. Where possible, I will be following the notational conventions used by Griffiths in his *Introduction to Electrodynamics* [1] for consistency with Regis's other upper division physics courses, and this review chapter draws heavily from that textbook.

2.1 Basic Vector Operations

Light is an electromagnetic wave, and both electric and magnetic fields are vectors, so we will need to know how vectors behave mathematically if we are to study light.

2.1.1 Vectors

A **vector** is a mathematical quantity that has both a magnitude and a direction; common examples of vectors in physics include velocity, acceleration, force, momentum, and displacement. Vectors are easiest to think of graphically. If we draw an arrow, it has both a magnitude (length) and a direction in which it is pointing, so arrows are typically used to represent vectors in diagrams. These properties disginguish vectors from **scalars**, which have a magnitude but no direction. In physics, common scalars include charge, temperature, mass, and density.

In type, vectors are generally notated via the use of an arrow above the variable (e.g., $\vec{E}$, $\vec{B}$) or via the use of boldface (e.g., $\mathbf{E}$, $\mathbf{B}$). If we wish to refer to the magnitude of a vector, which is a scalar quantity, we need an additional piece of notation to distinguish it from the vector itself. Usually, the magnitude of the vector $\mathbf{E}$ is written $|\mathbf{E}|$ or simply E.

So far, I've described vectors only using abstract symbols that do not reference any particular coordinate system. In practice, however, one often needs to be more specific about how a vector lies in a particular coordinate system. There are two main ways of doing so:

- **Specify magnitude and direction separately.** We could, for example, say that a vector has a magnitude of 4 m and it is pointed 20° north of east.
- **Specify the vector's components separately.** We could, for example, say that the same vector has a northward component of 1.37 m and an eastward component of 3.76 m. Put another way, to get from the vector's tail to its head, we would need to travel 1.37 m north and 3.76 m east.

Both approaches convey the same information in different formats, but notice that both required specifying a coordinate system (in this case, the cardinal directions on earth). We can convert between the two formats by using trigonometry, though the component form is usually most useful in physics.

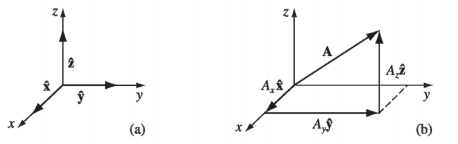

Figure 2.1 Unit vectors and vector components. Figure from [1].

If we are working in 3-dimensional space, we will most often set up an x, y, z Cartesian coordinate system. In such a coordinate system, each vector will have a component for each axis; vector $\mathbf{A}$ has components A_x, A_y, and A_z. We further define the **unit vectors** $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, shown in the left panel of Figure 2.1. Unit vectors are indicated with a "hat" symbol, have a magnitude of 1, point in the direction of the "hatted" vector, and have no units. Using unit vectors, we can express any vector in terms of its components:

$$\mathbf{A} = A_x\hat{\mathbf{x}} + A_y\hat{\mathbf{y}} + A_z\hat{\mathbf{z}}. \tag{2.1}$$

2.1.2 Vector Operations

We can define four common vector operations: addition and three kinds of multiplication:

1. **Addition.** If we place the tail of $\mathbf{B}$ at the head of $\mathbf{A}$ as in the left-most triangle of Figure 2.2, then the sum $\mathbf{A}+\mathbf{B}$ is the vector that points from the tail of $\mathbf{A}$ to the head of $\mathbf{B}$. Addition is commutative, so we find that $\mathbf{A}+\mathbf{B} = \mathbf{B}+\mathbf{A}$, as shown in the middle triangle of Figure 2.2. As with scalar addition, vector addition is also associative. If we wish to subtract a vector, we can simply add its opposite: $\mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B})$. Vector subtraction is illusrated in the right-most triangle of Figure 2.2.

   Vectors in component form can be added by adding like components:

$$\begin{aligned}\mathbf{A} + \mathbf{B} &= (A_x\hat{\mathbf{x}} + A_y\hat{\mathbf{y}} + A_z\hat{\mathbf{z}}) + (B_x\hat{\mathbf{x}} + B_y\hat{\mathbf{y}} + B_z\hat{\mathbf{z}}) &(2.2)\\ &= (A_x + B_x)\,\hat{\mathbf{x}} + (A_y + B_y)\,\hat{\mathbf{y}} + (A_z + B_z)\,\hat{\mathbf{z}} &(2.3)\end{aligned}$$

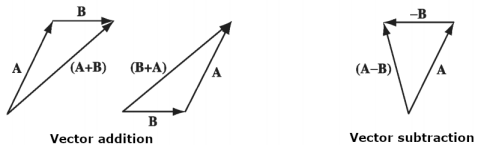

Figure 2.2 Graphical vector addition and subtraction. Figure adapted from [1].

2. **Multiplication by a scalar.** Multiplication of a vector by a positive scalar changes the magnitude of the vector but not the direction, as in the left panel of Figure 2.3. Such an operation can be carried out by multiplying each vector component by the scalar: $a\mathbf{A} = (aA_x)\,\hat{\mathbf{x}} + (aA_y)\,\hat{\mathbf{y}} + (aA_z)\,\hat{\mathbf{z}}$.

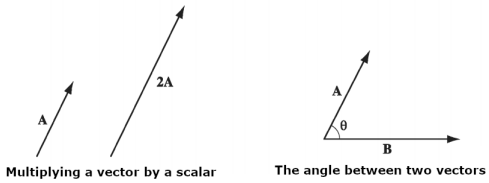

Figure 2.3 Graphical vector multiplication. Figure adapted from [1].

3. **Dot product of two vectors.** We define the dot prodcut of two vectors as

$$\mathbf{A} \cdot \mathbf{B} \equiv AB\cos\theta, \tag{2.4}$$

where θ is the angle between the two vectors when they are placed tail-to-tail, as shown in Figure 2.3. Notice that the dot product always yields a scalar. It is both commutative (i.e., $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$) and distributive (i.e., $\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C}$). In terms of components,

$$\mathbf{A} \cdot \mathbf{B} = A_x B_x + A_y B_y + A_z B_z. \tag{2.5}$$

Geometrically, $\mathbf{A} \cdot \mathbf{B}$ is the product of A times the projection of $\mathbf{B}$ along $\mathbf{A}$, so it tells us something about how much the two vectors lie in the same direction. If the vectors are exactly parallel, then the dot product is maximized, while if the two vectors are exactly perpendicular, the dot product is zero.

4. **Cross product of two vectors.** The cross product is defined as

$$\mathbf{A} \times \mathbf{B} \equiv AB\sin\theta\ \hat{\mathbf{n}}, \tag{2.6}$$

where $\hat{\mathbf{n}}$ is a unit vector pointing perpendicular to both $\mathbf{A}$ and $\mathbf{B}$ in a way that agrees with the right-hand rule. Notice that, unlike the dot product, the cross product is a vector. The cross product, like the dot product, is distributive, but, unlike the dot product, it is not commutative. If two vectors are parallel, their cross product is zero, and if they are perpendicular, then the magnitude of the cross product is maximized.

Therefore, the cross product is in some sense a measure of how perpendicular two vectors are. The cross product can be calculated using components, but the expression is quite messy:

$$\mathbf{A} \times \mathbf{B} = (A_y B_z - A_z B_y)\,\hat{\mathbf{x}} + (A_z B_x - A_x B_z)\,\hat{\mathbf{y}} + (A_x B_y - A_y B_x)\,\hat{\mathbf{z}}. \tag{2.7}$$

If you are familiar with the concept of a determininant from a linear algebra course, the cross product is more easily memorized and written as a determinant:

$$\mathbf{A} \times \mathbf{B} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix}. \tag{2.8}$$

All of these operations can also be performed in other coordinate systems, such as polar coordinates, but those formulae are beyond the scope of this chapter.

2.2 Derivatives and the Del Operator

2.2.1 Ordinary and Partial Derivatives

If we have a function of one variable, $f(x)$, then we can calculate its derivative df/dx, which tells us how rapidly the function varies. To be preceise, when we change x by a tiny amount dx, the derivative allows us to determine the tiny amount by which f changes:

$$df = \left(\frac{df}{dx}\right) dx. \tag{2.9}$$

This means the the derivative df/dx is the slope of the graph of f versus x.

But what if we have a function of three variables, such as temperature as a function of position in a three-dimensional room, $T(x, y, z)$? We might want to know by what amount dT the function T changes if we alter x, y, and z. As is proved in any calculus textbook, we can generalize Equation 2.9 to three dimensions as

$$dT = \left(\frac{\partial T}{\partial x}\right) dx + \left(\frac{\partial T}{\partial y}\right) dy + \left(\frac{\partial T}{\partial z}\right) dz. \tag{2.10}$$

Equation 2.10 contains three **partial derivatives**, which each tells us how T responds to changes in just one variable. For example, if $T = 3xy + z$, we can compute $\partial T/\partial x = 3y$ by simply treating y and z as constants while taking a derivative with respect to x.

If we stare long enough and squint hard enough, we might notice that Equation 2.10 has a form quite similar to a dot product, and indeed, we can write it as one:

$$dT = \left(\frac{\partial T}{\partial x}\hat{\mathbf{x}} + \frac{\partial T}{\partial y}\hat{\mathbf{y}} + \frac{\partial T}{\partial z}\hat{\mathbf{z}}\right) \cdot \left(dx\;\hat{\mathbf{x}} + dy\;\hat{\mathbf{y}} + dz\;\hat{\mathbf{x}}\right). \tag{2.11}$$

This equation becomes easier to intepret if we define two additional quantities:

$$d\mathbf{l} \equiv dx\ \hat{\mathbf{x}} + dy\ \hat{\mathbf{y}} + dz\ \hat{\mathbf{x}} \quad \text{and} \quad \nabla T \equiv= \frac{\partial T}{\partial x}\hat{\mathbf{x}} + \frac{\partial T}{\partial y}\hat{\mathbf{y}} + \frac{\partial T}{\partial z}\hat{\mathbf{z}}, \tag{2.12}$$

where $d\mathbf{l}$ is simply a displacement and ∇T is the **gradient** of T. With these definitions, Equation 2.11 becomes

$$dT = \nabla T \cdot d\mathbf{l} = |\nabla T|\,|d\mathbf{l}| \cos\theta. \tag{2.13}$$

Comparison to Equation 2.9 suggests that ∇T plays a similar role as df/dx did in Equation 2.9. Further, for some fixed-magnitude $d\mathbf{l}$, it seems that dT is biggest when $\theta = 0$, i.e., when $d\mathbf{l}$ and ∇T point in the same direction. This all suggests a geometrical interpretation: the gradient ∇T, which is a vector, points in the direction of maximum increase of the function T, and the magnitude $|\nabla T|$ gives the slope along this maximal direction.

This geometrical interpretation is easier to visualize in two dimensions rather than in three. For example, imagine you are standing on a hill on earth. Your elevation only depends on two position coordinates, latitude and longitude. If you look all around you, you can find the direction of steepest ascent. This is the direction of the gradient at your position. If you measure the slope in that direction, then you have found the magnitude of the gradient at your location.

2.2.2 The Del Operator

My goal with this discussion of derivatives so far has been to arrive at a definition of ∇, usually known as the **del operator**, del, or nabla. If you recall Maxwell's Equations, which govern electromagnetism and thus light, you might recall that ∇ sometimes appears within them. If we wish to start from Maxwell's Equations in our study of light, then we need to make sure that we can make sense of how they are written down.

A careful inspection of the definition of ∇T in Equation 2.12 reveals how we should think about the del operator. A gradient seems to have the appearance of a "vector," ∇, multiplying a scalar, T, so it seems that we could write it as

$$\nabla T = \left(\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right) T. \tag{2.14}$$

In this form, it is obvious that the term in parentheses must be the definition of ∇:

$$\nabla = \hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}. \tag{2.15}$$

It's tempting to look at this and say that ∇ *is* a vector. But that can't be right – this quantity has no value at all unless it has some function to operate upon. It's also clear that the gradient *is not* simply a multiplication; it's a calculus operation that just happens to be written in a form that looks like multiplication. So, really, ∇ is a **vector operator**:

it doesn't mean anything unless we give it a function to operate upon. However, thinking about its resemblance to a vector is an extremely handy notational trick, because if we merely mentally translate the word "multiply" into "acts upon," everything about ∇ behaves like a vector.

If we think about our basic vector operations, we can multiply a vector in three ways: by a scalar, by a vector via the dot product, or by a vector via the cross product. Similarly, there are three ways that the operator ∇ can act:

1. On a scalar function T: ∇T (the gradient);
2. On a **vector function**, $\mathbf{v}$, via the dot product: $\nabla \cdot \mathbf{v}$ (the **divergence**);
3. On a **vector function**, $\mathbf{v}$, via the cross product: $\nabla \times \mathbf{v}$ (the **curl**).

We've already discussed the gradient, but those other two operations introduce three new pieces of jargon, so let's take them one by one.

Vector Functions

It's clear that ∇, because it contains derivatives, can't usefully act upon a single number; we take derivatives of functions. When we took the gradient, we applied those derivatives to a scalar function, which is just a function T that for every combination of x, y, and z, $T(x, y, z)$ spits out a scalar value.

But dot products and cross products aren't multiplications of scalars; they are multiplications of two vectors. We therefore need a vector function for the ∇ to act upon. A vector function is really a combination of three scalar functions that play the role of the three vector components:

$$\mathbf{v}(x, y, z) = v_x(x, y, z)\hat{\mathbf{x}} + v_y(x, y, z)\hat{\mathbf{y}} + v_z(x, y, z)\hat{\mathbf{z}}. \tag{2.16}$$

Thus, if we give $\mathbf{v}(x, y, z)$ a combination of x, y, and z as input, $\mathbf{v}(x, y, z)$ spits out a vector. It returns a different vector for every possible combination of x, y, and z.

The Divergence

From the definitions of ∇ and the dot product, we can construct the divergence:

$$\nabla \cdot \mathbf{v} = \left(\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right) \cdot (v_x\hat{\mathbf{x}} + v_y\hat{\mathbf{y}} + v_z\hat{\mathbf{z}}) = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}. \tag{2.17}$$

Notice that even though $\mathbf{v}$ is a vector function, it's divergence produces a scalar, as we would expect from a dot product. But what does the divergence mean? The name is revealing: the divergence is a measure of how much the vector $\mathbf{v}$ spreads out (diverges) from the point in questtion, as is explored much more fully in our electromagnetism course.

The Curl

From the definitions of ∇ and the dot product, we can construct the curl:

$$\nabla \times \mathbf{v} = \left(\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right) \times (v_x\hat{\mathbf{x}} + v_y\hat{\mathbf{y}} + v_z\hat{\mathbf{z}}) \tag{2.18}$$

$$= \hat{\mathbf{x}}\left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right) + \hat{\mathbf{y}}\left(\frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x}\right) + \hat{\mathbf{z}}\left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right). \tag{2.19}$$

Notice that like any cross product, the curl is a vector. Just as divergence was well named, so is curl. The curl is a measure of how much a vector $\mathbf{v}$ swirls around a point in question. Again, the nuances of curls are discussed extensively in our electromagnetism course.

2.2.3 Anything else we need?

We're now almost in a position to talk about the wave equation, what it has to do with light, and how that tells us about optics. The one remaining thing that we could do is construct second derivatives from these operations. For example, there's no reason that we couldn't take the divergence of a gradient (which is known as the Laplacian for short) or the curl of a curl. Doing so is nothing more than further algebra with dot products and cross products, so I won't spell out the results here – it's not our goal for this course. However, we *will* need these results. In particular, we will need the following two results:

1. The divergence of a gradient is known as the Laplacian:

$$\nabla \cdot (\nabla T) = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \equiv \nabla^2 T. \tag{2.20}$$

   Notice that the Laplacian of a scalar T is a scalar. However, we will sometimes use a notation that seems to indicate taking the Laplacian of a vector: $\nabla^2\mathbf{v}$. By this, we are using shorthand notation to indicate the following vector:

$$\nabla^2\mathbf{v} \equiv \left(\nabla^2 v_x\right)\hat{\mathbf{x}} + \left(\nabla^2 v_y\right)\hat{\mathbf{y}} + \left(\nabla^2 v_z\right)\hat{\mathbf{z}}. \tag{2.21}$$

2. The curl of the curl is

$$\nabla \times (\nabla \times \mathbf{v}) = \nabla\left(\nabla \cdot \mathbf{v}\right) - \nabla^2\mathbf{v}. \tag{2.22}$$

We will need these two results soon. The first is easy to prove, but the second is a bit more tedious, so we'll just take it as a given (though I encourage you to try it!). With all that said, let's move on to waves.

Chapter 3

Waves

We all know that light is a wave – nearly anyone on the street could tell you that. But if you pressed someone to be a bit more specific on that final word –"wave" – most people would struggle to define it. Sure, it's easy to give examples of waves, but what, specifically, is a wave?

I'm not sure that I can give a concise definition in English words that is perfectly precise while still including all possible waves. Griffiths [1], upon whose work I am basing this section, says that a **wave** "is a disturbance of a continuous medium that propagates with a fixed shape at constant velocity." Of course, he's making some simplifying assumptions here, so this isn't completely general. He's assuming that there's no absorption (which would gradually dimish the wave), that the medium isn't dispersive (which would cause different frequencies to move at different speeds), that we're in one-dimension (because in 3D, the wave would spread out and weaken), and more. But let's stick with this simplified definition for now, which is illustrated in Figure 3.1. We might generate such a wave by doing all sorts of things; for example, we might shake the end of a taut string. It turns out that this simple conception of a wave will be sufficient to capture most phenomena that we are interested in at the moment.

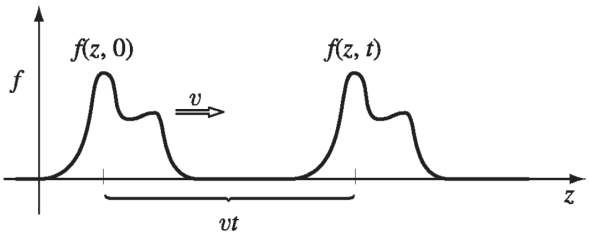

Figure 3.1 A simple wave. Figure from [1].

3.1 The Wave Equation

We need to represent the idea of a wave mathematically. Figure 3.1 shows a wave at two different times, once at $t = 0$ and again at a later time t. Each point on on the wave form shifts to the right by an amount vt, where v is the wave's velocity. If we imagine that this is a wave on a string, then the function $f(z, t)$ is the displacement of the string at the point z

at time t. Let's define the *initial* shape of the string to be $g(z)$, or, to put it mathematically, $g(z) \equiv f(z,0)$. If that's the case, can we express the later form $f(z,t)$ in terms of g? The displacement at point z, at the later time t, is the same as the displacement vt to the left (i.e., at $z - vt$) when the time was $t = 0$:

$$f(z,t) = f(z - vt, 0) = g(z - vt). \tag{3.1}$$

That equation captures the idea of wave motion – be sure you understand it! It says that the function $f(z,t)$ – which could depend on z and t in *any* way – in fact *only* depends on them in the special combination $z - vt$. When that is true, the function $f(z,t)$ represents a wave of fixed shape traveling in the z direction at speed v.

It might help to give a concrete example. Let's assume that A and b are constants of the appropriate units. In that case, all of the following are waves (of different shapes):

$$f_1(z,t) = Ae^{-b(z-vt)^2}, \qquad f_2(z,t) = A\sin\left[b\left(z - vt\right)\right], \qquad f_3(z,t) = \frac{A}{b\left(z - vt\right)^2 + 1}. \tag{3.2}$$

However, the following equations are *not* waves:

$$f_4(z,t) = Ae^{-b\left(bz^2 + vt\right)}, \qquad f_5(z,t) = A\sin(bz)\cos(bvt)^3. \tag{3.3}$$

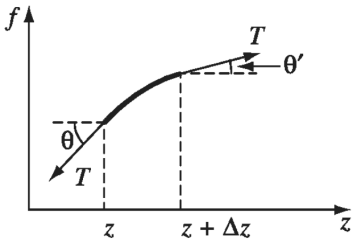

Figure 3.2 A taut string segment. Figure from [1].

So far, we have mathematically *described* a wave, but that doesn't really physically tell us why, physically, they occur, nor have we characterized that underlying process mathematically. Let's stick with our taut string example, so that we have something concrete to work with. In that case, if waves exist (and they do), it should follow from Newton's Second Law, like all mechanical processes. Imagine a very long string with tension T. If we displace a point on the string away from its equilibrium, then the net transverse (vertical in Figure 3.2) force on a small segment bewteen locations z and $z + \Delta z$ is

$$\Delta F = T\sin\theta' - T\sin\theta, \tag{3.4}$$

where θ' is the angle the string makes with the z-direction at point $z + \Delta z$, and θ is the corresponding angle at point z.

If we're imagining that the displacement of the string vertically and Δz are very small, then Figure 3.2 is a vast exaggeration – the angles will by *tiny*. When θ is small, i.e. $\theta \ll 1$, you can show via a Taylor series that both

$$\sin\theta \approx \theta \qquad \text{and} \qquad \tan\theta \approx \theta. \tag{3.5}$$

Therefore, there's no reason that we can't simply replace the sines with tangents in our force equation:

$$\Delta F \approx T\left(\tan\theta' - \tan\theta\right). \tag{3.6}$$

Why did I do that? Well, look at the geometry indicated in Figure 3.3. Remembering the definition of tangent, we can see that $\tan\theta = \Delta f/\Delta z \approx \partial f/\partial z$. Since this is a zoom-in of a tiny region, we expect the same derivative to be approximately valid at both z and $z + \Delta z$, so we can rewrite our force equation as

$$\Delta F \approx T\left(\left.\frac{\partial f}{\partial z}\right|_{z+\Delta z} - \left.\frac{\partial f}{\partial z}\right|_{z}\right). \tag{3.7}$$

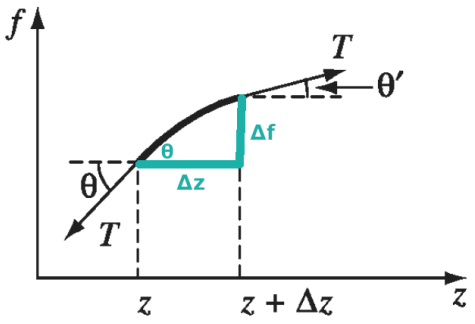

Figure 3.3 A taut string segment. Figure adapted from [1].

But that's starting to look a lot like the basic definition of a derivative. Indeed, one way of writing the definition of a second derivative is

$$\frac{d^2 f}{dz^2} = \lim_{dz\to 0} \frac{1}{dz}\left(\left.\frac{df}{dz}\right|_{z+dz} - \left.\frac{df}{dz}\right|_{z}\right), \tag{3.8}$$

so, because we are looking at a tiny region that does satisfy that limit, we can write our force equation as

$$\Delta F \approx T\,\frac{\partial^2 f}{\partial z^2}\,\Delta z. \tag{3.9}$$

Let's call the force per unit length μ, so that the string segment's mass is $\mu\,(\Delta z)$. Recalling that a second derivative of position with respect to time is an acceleration, according to Newton's Second Law,

$$\Delta F = \mu\,(\Delta z)\,\frac{\partial^2 f}{\partial t^2}. \tag{3.10}$$

Setting Equations 3.9 and 3.10 equal to each other, we find that

$$\frac{\partial^2 f}{\partial z^2} = \frac{\mu}{T}\frac{\partial^2 f}{\partial t^2}. \tag{3.11}$$

Thinking about the units of T and μ reveals that their combination defines a velocity, $v = \sqrt{T/\mu}$, so our result can also be written as

$$\frac{\partial^2 f}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2}. \tag{3.12}$$

Evidently, small disturbances on our string satisfy Equation 3.12, and correspondingly, that equation is usually referred to as the **wave equation**. Though we imagined a disturbance on a string while deriving it so that we would have something concrete to talk about, we didn't actually make any assumptions that only apply to strings. All of our steps could apply to any physical medium, so this equation must be generally true for small disturbances in any medium.

This differential equation is called the wave equation because it is satisfied by all functions of the form

$$f(z,t) = g(z - vt), \tag{3.13}$$

which we have previously reasoned out describes what we think of as a wave. The proof is fairly straightforward, so I encourage you to check my claim. Do so by setting $u \equiv z - vt$, and then rewriting the derivatives of f in terms of g and u via a change of variables. You should ultimately find that

$$\frac{\partial^2 g}{\partial u^2} = \frac{\partial^2 f}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2}, \tag{3.14}$$

verifying my claim.

Perhaps unsurprisingly, functions of the form $g(z - vt)$ aren't the *only* functions that satisfy the wave equation. It is also satisfied by equations of the form $h(z + vt)$, which simply correspond to waves moving in the negative z direction. The most general solution to the wave equation is thus any linear combination of waves moving left and right:

$$f(z,t) = g(z - vt) + h(z + vt). \tag{3.15}$$

The wave equation appears frequently in most branches of physics and science more generally, because it, in various forms, describes anything that has a wave-like behavior. For our purposes, it will reveal properties of optics, but you are just as likely to encounter it while studying oceanography, musical instruments, or the engineering of bridges.

3.2 Sinusoidal Waves

We can further restrict our discussion to the familiar sinusoidal (or, equivalently, cosinusoidal) wave form. On the surface, this might seem like a really restrictive approach, because our definition of waves included infinitely more possible waveforms, but we'll see that that doesn't turn out to be a problem.

3.2.1 Defining the Sinusoidal Wave

When we speak of a sinusoidal wave, we are refering to a function of the form

$$f(z,t) = A\cos\left[k\left(z - vt\right) + \delta\right] \tag{3.16}$$

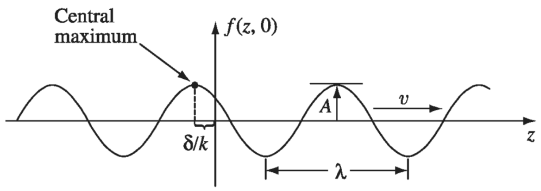

Figure 3.4 A sine wave. Figure from [1].

Figure 3.4 plots this function at time $t = 0$. Let's define some jargon for the different pieces of this equation. In my experience, students tend to have poor intuition for how variables affect trigonometric functions when they first start using such functions extensively in physics. It's worth taking time to go through this list to make sure that you understand how and why each of these variables affects f when they are changed. Check yourself in this mental exercise by using an online plotting tool.

- The constant A is the **amplitude**. It is positive and represents the maximum displacement from equilibrium.
- The entire argument of the cosine is called the **phase**.
- The unitless constant δ is called the **phase constant**. You could add any integer multiple of 2π to δ without changing the value of $f(z, t)$. However, usually one would just use a value in the range $0 \leq \delta < 2\pi$ for simplicity.
- At $z = vt - \delta/k$, the phase is zero, so the cosine is maximized. Some authors thus call the point at $z = vt - \delta/k$ the **central maximum** (though this piece of jargon is not nearly as ubiquitous as the others in this list). If $\delta = 0$, notice that the central maximum passes the origin at $t = 0$. This suggests a useful interpretation: δ/k is the distance by which the central maximum (and the wave as a whole) is "delayed" compare to a cosine with zero phase constant.
- The constant k is known as the **wave number**, because it is closely related to the **wavelength** λ (the distance between wave peaks). When z advances by $2\pi/k$, the cosine completes one full cycle, so it must be that $\lambda = 2\pi/|k|$.
- As time passes, the entire wave proceeds to the right at speed v, so any single point on the string will oscillate up and down. We call the time to complete one full cycle a **period**, and according to our definition of wavelength, we can calculate the period as $T = 2\pi/kv$.
- The reciprocal of the period, $\nu = T^{-1} = v/\lambda$, is called the **frequency**, which tells us the number of oscillations that occur per unit time.
- Finally, a more convenient unit than frequency is the **angular frequency**: $\omega = 2\pi\nu = kv$. It gets its name because in the analogous case of uniform circular motion, which is also described by a sinusoidal function, it represents the number of radians swept out per unit time.

For our purposes, it is much cleaner to write our waveform in terms of ω rather than v:

$$f(z,t) = A\cos\left(kz - \omega t + \delta\right). \tag{3.17}$$

As we've previously seen, we can flip signs to create an identical wave that moves left instead of right:

$$f(z,t) = A\cos\left(kz + \omega t - \delta\right). \tag{3.18}$$

In this equation, I've also flipped the sign of δ. Of course, there are infinite equally-valid choices for δ, but this sign convention maintains our previous convention of thinking of δ/k as the distance by which the wave is "delayed." Since the wave is now moving left, a delay means a shift to the right, so it makes sense to flip the sign of δ for clarity. At $t = 0$, the wave described by Equation 3.18 is identical to the one shown in Figure 3.4, but at later times it becomes apparent that it is moving in the opposite direction. Because the cosine is an even function, we could, if we wanted, choose to write Equation 3.18 as

$$f(z,t) = A\cos\left(-kz - \omega t + \delta\right). \tag{3.19}$$

This form is very revealing: comparison to Equation 3.17 shows that we can simply switch the sign of k to produce a wave traveling in the opposite direction. In 1D, the sign of the constant k thus tells us something about the direction in which a wave is propagating. Analogously, when we generalize this equation to 3D, we will see that k becomes a vector, and its direction indicates the direction of propagation.

3.2.2 Waves in Complex Notation

As we all likely learned (and then, perhaps, forgot) in some math class from our past, Euler's formula relates sines and cosines to exponential functions via the imaginary number:

$$e^{i\theta} = \cos\theta + i\sin\theta. \tag{3.20}$$

We can therefore rewrite our sinusoidal wave (Equation 3.17) as

$$f(z,t) = \mathrm{Re}\left[Ae^{i(kz - wt + \delta)}\right], \tag{3.21}$$

where the notation $\mathrm{Re}\,(\xi)$ simply indicates the real part of the complex number ξ. To simplify our notation, we will often just talk about the complex wave function,

$$\tilde{f}(z,t) \equiv \tilde{A}e^{i(kz - wt)}, \tag{3.22}$$

with the complex amplitude, $\tilde{A} \equiv Ae^{i\delta}$, absorbing the phase constant for convenience. Of course, the *actual* wave function is just the real part of $\tilde{f}$:

$$f(z,t) = \mathrm{Re}\left[\tilde{f}(z,t)\right]. \tag{3.23}$$

Initially, this additional notation might seem to pointlessly add complexity. However, exponentials are *far* easier to manipulate than trigonometric functions, and if we know $\tilde{f}$ then it is easy to find f, so this notation will make almost all math involving waves much easier. If you try adding together two waves to determine the amplitude and phase constant of the result, you'll see what I mean: you'll get your answer almost immediately if you do it with exponentials, but you'll probably have to dig for obscure trig identities online if you don't.

3.2.3 Linear Combinations of Sinusoidal Waves

All this work has only applied to sinusoidal waves, but that's OK, because sinusoidal waves are a special class of waves. It turns out that *any* wave at all can be espressed as a linear combination of sinusoidal ones (though it might take an infinite number of them). Keeping in mind that w is a function of k, we can express this as

$$\tilde{f}(z,t) = \int_{-\infty}^{\infty} \tilde{A}(k) e^{i(kz-wt)} dk. \tag{3.24}$$

As a reminder, here we are using $\tilde{f}$ to refer to any possible (complex) wave that we might want to consider, so this equation is saying that any wave can be constructed by some potentially infinite combination sine waves with different frequencies and amplitudes. Adding up all of those sine waves (i.e., doing the integral) constructs the wave of interest, $\tilde{f}$.

The business of actually figuring out what $\tilde{A}(k)$ should be depends on the conditions of f and its derivative at $t = 0$, and the procedure requires knowledge of Fourier transforms. We'll encounter Fourier transforms later in the semester. For now, the details aren't important. What's important is just to realize that by studying sinusoidal waves, we are actually studying all possible waves!

Chapter 4

Electromagnetic Waves

We've covered the basic math of waves in general, but how does this apply to optics? We could very quickly derive the laws of refraction and reflection using hypotheses known as Fermat's Principle or Huygen's Principle (and these are very interesting derivations – google them!), but these approaches kind of side-step how all of this relates to the most fundamental laws of physics that you study in other physics classes. So, instead, let's take a look at Maxwell's Equations, and see if we can reason out our optical laws directly from those more fundamental physical laws. I'll again be following Griffiths [1] in terms of notation and approach, for consistency with our other upper division physics classes.

4.1 The Wave Equation for E and B

Let's talk about some homogeneous region of space. It might not be vacuum, but let's assume that there is **no** *free* charge or current. That is, there might be atoms that comprise some medium, but that medium is neutral, homogenous, and has no bulk electrical current running through it. Let's further assume that any material that is present is *linear*. This is an approximation, but it is a fairly general one – most materials are roughly linear. If you've taken upper division electromagnetism, this simply means that we are assuming that we can write the **D** and **H** fields as

$$\mathbf{D} = \epsilon \mathbf{E} \qquad \text{and} \qquad \mathbf{H} = \frac{1}{\mu}\mathbf{B}. \tag{4.1}$$

If you haven't taken upper-division electromagnetism, and have no idea what **D** and **H** fields are, don't worry. Basically, we're just assuming that the degree to which our material affects electromagnetic fields can be captured by two simple constants, ϵ and μ, which are known as the permittivity and permeability of the material, respectively. You might recall from Physics 2 that there are very similar-looking constants for electromagnetism in a vacuum: ϵ_0 and μ_0. Our two new (non-subscripted) constants play the same role as those familiar constants, except they are material-specific rather than only applicable in a vacuum.

Our upper division electromagnetism course derives Maxwell's equations, and our Physics 2 class discusses them qualitatively, so we will simply state them in this course as tools that we can use. In a homogeneous, linear medium with no free charge or free current, Maxwell's equations can be written as:

$$\nabla \cdot \mathbf{E} = 0 \qquad \text{(Gauss's Law)} \tag{4.2}$$
$$\nabla \cdot \mathbf{B} = 0 \qquad \text{(no magnetic monopoles)} \tag{4.3}$$
$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{(Faraday's Law)} \tag{4.4}$$
$$\nabla \times \mathbf{B} = \mu\epsilon\frac{\partial \mathbf{E}}{\partial t} \qquad \text{(Ampere's Law w/ Maxwell's Correction)} \tag{4.5}$$

Though Maxwell's equations are incredibly elegant in some sense, they also can generate some pretty messy math. As written above, they are a set of first-order partial differential vector equations. They are also *coupled*, i.e., the electric and magnetic fields appear in the same equations together, so we can't treat them completely independently. However, we can *decouple* them by taking the curls of Faraday's and Ampere's Laws. Taking the curl of both sides of Equation 4.4 and using Equation 2.22, we find that

$$\nabla\left(\nabla \cdot \mathbf{E}\right) - \nabla^2\mathbf{E} = \nabla \times \left(-\frac{\partial \mathbf{B}}{\partial t}\right). \tag{4.6}$$

Substituting Gauss's Law into the left-hand side and rearranging the derivatives on the right-hand side gives

$$\nabla^2\mathbf{E} = \frac{\partial}{\partial t}\left(\nabla \times \mathbf{B}\right). \tag{4.7}$$

Substituting Ampere's Law into the right side gives

$$\nabla^2\mathbf{E} = \mu\epsilon\frac{\partial^2 \mathbf{E}}{\partial t^2}. \tag{4.8}$$

You can (and you should, right now!) use a similar approach to show that

$$\nabla^2\mathbf{B} = \mu\epsilon\frac{\partial^2 \mathbf{B}}{\partial t^2}. \tag{4.9}$$

Now we have separate equations for the two fields, but we had to pay a price: now they have second derivatives in them. But wait! Compare these last two equations to the wave equation we derived in Equation 3.12, and think about the definition of the vector Laplacian. Each of these equations is really just a vector comprised of three 3D wave equations. That is, each Catesian component of these vector equations is a scalar 3D wave equation of the form

$$\nabla^2 f = \mu\epsilon\frac{\partial^2 f}{\partial t^2}. \tag{4.10}$$

And *this* form is just the three-dimensional generalization of the one-dimensionalwave equation; our 1D spatial second derivative is simply replaced by its 3D cousin, the Laplacian. The 3D wave equation is solved by any wave-like function in 3D space.

So evidently, electromagnetic waves propagate through a material at a speed $v = 1/\sqrt{\epsilon\mu}$. If we use the familiar vacuum values for the constants, then we have derived the familiar speed

of light in a vacuum, c. Cool! Often, the speed of light in a material is convenient to express in terms of the speed of light in a vacuum: $v = c/n$, where

$$n \equiv \sqrt{\frac{\epsilon\mu}{\epsilon_0\mu_0}} \tag{4.11}$$

is the familiar **index of refraction**. The index of refraction is simply a dimensionless number that tells us how many times faster the speed of light is in a vacuum compared to the material in question (because $c = nv$).

Given the wave equations we found, the most generic forms of monochromatic (i.e., single frequency) sinusoidal electromagnetic waves in one dimension must be

$$\tilde{\mathbf{E}}(z,t) = \tilde{\mathbf{E}}_0 e^{i(kz-\omega t)} \quad \text{and} \quad \tilde{\mathbf{B}}(z,t) = \tilde{\mathbf{B}}_0 e^{i(kz-\omega t)}. \tag{4.12}$$

Of course, these waves, as written, are complex; only the real part is the actual physical wave. There's a lot more to say about these waves, but most of it isn't directly relevant to our purposes at the moment, so I will not derive all of their properties. In particular, it's easy to prove that these waves are transverse, i.e., they oscillate perpendicular to the direction of propagation. Moreover, the electric and magnetic waves are related, as you should have seen in Physics 2: they are always perpendicular to each other, and their amplitudes are related by just a constant. Finally, these two waves are always in phase.

Of course, there is nothing special about the z direction. We can generalize these expressions to characterize waves traveling in an arbitrary direction in 3D space. To do so, we define $\mathbf{k}$, known as either the **propagation vector** or **wave vector**. The vector $\mathbf{k}$ points in the direction of propagation, and its magnitude is the wave number k. Therefore $\mathbf{k}$, $\mathbf{E}$, and $\mathbf{B}$ are all perpendicular to each other. Let's further define the position vector $\mathbf{r}$, which is simply the position vector of some point in space relative to the origin (i.e., it is the 3D analogue of z in our familiar 1D wave). Our wave functions then become

$$\tilde{\mathbf{E}}(\mathbf{r},t) = \tilde{\mathbf{E}}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \quad \text{and} \quad \tilde{\mathbf{B}}(\mathbf{r},t) = \tilde{\mathbf{B}}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}. \tag{4.13}$$

These waves, of course, represent only one possible freqeuency, but, remembering what I said about linear combinations of sine waves, we could always construct any pattern of light out of these building blocks.

4.2 The Electrodynamic Boundary Conditions

From an optics point of view, the interesting question is this: what happens when an electromagnetic wave passes from one medium into another? Or reflects off a medium? If we can answer those questions, we will have a set of rules that let us design optical devices, i.e., devices that control the behavior of light. These questions ask what happens at a boundary between two materials; we need to know the **boundary conditions** that govern electromagnetic waves at such an interface. Put another way, consider the situation in Figure 4.1. The

centeral black line is the boundary between two materials, which have different permittivities and permeabilities. We want to know if there are any rules that govern how the electric and magnetic fields of a wave just barely to the left of boundary relate to the fields of that way just barely on the other side of the boundary. Are they continuous across the boundary? Discontinuous? Something more complicated?

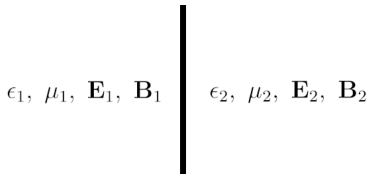

Figure 4.1 A boundary between two linear materials.

This question is one of the major questions of our upper-division electromagnetism class, where all of the results are derived in detail. Here, we will simply state the results, discuss them qualitatively, and use them. Let's make our notation very clear. As we have so far, we are still assuming that there is no free charge or current present, and that the materials are linear and homogeneous. In this section, the subscripts 1 and 2 will refer to properties at locations just barely to left and just barely to the right of the boundary between the two linear materials, respectively. Because electric and magnetic fields are vectors, we can split them up into components when we talk about them. Let's use the symbol $\parallel$ to indicate components parallel to the boundary (i.e., components going into, out of, up, or down in Figure 4.1 at the location of the boundary). Similarly, we'll use the $\perp$ symbol to indicate the components that are perpendicular to the boundary surface. With these notation conventions, we can list the boundary conditions, which are derived in our upper-division electromagnetism class or any electrodynamics textbook.

The tangential components of the electric field are continuous across the boundary:

$$\mathbf{E}_1^{\parallel} = \mathbf{E}_2^{\parallel}. \tag{4.14}$$

Similarly, the perpendicular components of the magnetic field are continuous across the boundary:

$$B_1^{\perp} = B_2^{\perp}. \tag{4.15}$$

In constrast, the perpendicular components of the electric field are discontinuous across the boundary, and the size of the discontinuity depends of the permittivities of the materials:

$$\epsilon_1 E_1^{\perp} = \epsilon_2 E_2^{\perp}. \tag{4.16}$$

Similarly, the parallel component of the magnetic field is discontinuous in proportion to the permeabilities:

$$\frac{1}{\mu_1}\mathbf{B}_1^{\parallel} = \frac{1}{\mu_2}\mathbf{B}_2^{\parallel}. \tag{4.17}$$

Remarkably, our knowledge of waves combined with these boundary conditions will be enough to deduce the laws that govern the reflection and refraction of light.

4.3 The Reflection and Transmission of Light

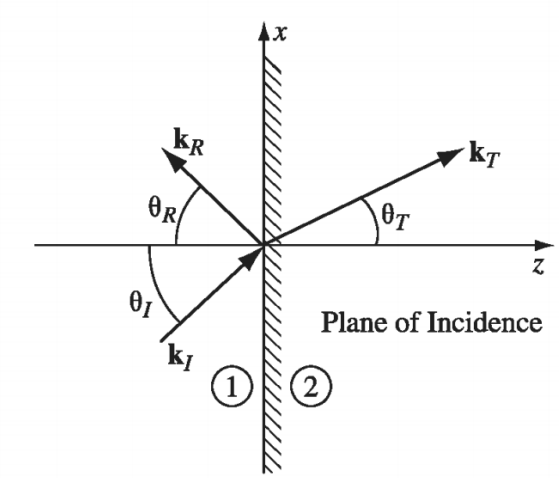

Figure 4.2 Incident, reflected, and transmitted waves at a boundary. Figure from [1].

Suppose we have an incoming electromagnetic wave moving approaching a boundary from the left, as in Figure 4.2. We can express this wave as

$$\tilde{\mathbf{E}}_I(\mathbf{r},t) = \tilde{\mathbf{E}}_{0_I} e^{i(\mathbf{k}_I\cdot\mathbf{r}-\omega t)} \quad \text{and} \quad \tilde{\mathbf{B}}_I(\mathbf{r},t) = \tilde{\mathbf{B}}_{0_I} e^{i(\mathbf{k}_I\cdot\mathbf{r}-\omega t)}. \tag{4.18}$$

This incoming wave is incident upon the boundary at some angle θ_I with respect to the normal. Material 1 is on the left of the boundary, and material 2 is on the right. Upon encountering the boundery, our incident wave could give rise to a reflected wave,

$$\tilde{\mathbf{E}}_R(\mathbf{r},t) = \tilde{\mathbf{E}}_{0_R} e^{i(\mathbf{k}_R\cdot\mathbf{r}-\omega t)} \quad \text{and} \quad \tilde{\mathbf{B}}_R(\mathbf{r},t) = \tilde{\mathbf{B}}_{0_R} e^{i(\mathbf{k}_R\cdot\mathbf{r}-\omega t)}, \tag{4.19}$$

and a transmitted wave,

$$\tilde{\mathbf{E}}_T(\mathbf{r},t) = \tilde{\mathbf{E}}_{0_T} e^{i(\mathbf{k}_T\cdot\mathbf{r}-\omega t)} \quad \text{and} \quad \tilde{\mathbf{B}}_T(\mathbf{r},t) = \tilde{\mathbf{B}}_{0_T} e^{i(\mathbf{k}_T\cdot\mathbf{r}-\omega t)}. \tag{4.20}$$

I've written all three waves as having the same freqeuncy ω, as *must* be the case for linear media. The frequency was set by whatever oscillated to create the incident wave in the first place (perhaps electrons moving in an antenna, for example), and nothing has been done to change that initial frequency. However, the wave numbers and velocities need not be the same for all three waves. These quantities are related by our original definition of angular frequency:

$$k_I v_1 = k_R v_1 = k_T v_2 = \omega, \quad \text{or} \quad k_I = k_R = \frac{v_2}{v_1} k_T = \frac{n_1}{n_2} k_T. \tag{4.21}$$

Remember from Physics 2 that both electric and magnetic fields obey superposition; the total field in a region is simply the sum of all the contributing fields in that region. Thus the total, combined fields in medium 1 in Figure 4.2 must be $\tilde{\mathbf{E}}_I + \tilde{\mathbf{E}}_R$ and $\tilde{\mathbf{B}}_I + \tilde{\mathbf{B}}_R$, while in medium

2 the total fields are simply $\tilde{\mathbf{E}}_T$ and $\tilde{\mathbf{B}}_T$. The two sets of total fields meet at the interface between the two materials, so they must be subject to the four boundary conditions listed in Section 4.2. We could write down an equation for each of those four boundary conditions. Each of these equations would have the same basic form, though the constants out in front of each exponential would be different. Mathematically, each boundary condition will yield an equation of the form

$$(\ )\, e^{i(\mathbf{k}_I \cdot \mathbf{r} - \omega t)} + (\ )\, e^{i(\mathbf{k}_R \cdot \mathbf{r} - \omega t)} = (\ )\, e^{i(\mathbf{k}_T \cdot \mathbf{r} - \omega t)} \quad \text{at } z = 0, \tag{4.22}$$

where the stuff that would go in the parentheses depends on which boundary condition we apply. Think extremely carefully about what I'm saying in Equation 4.22. It applies at any location where $z = 0$. The stuff in parentheses, whatever it is, can't possibly depend on x, y, or t, because each of those variables only shows up in the exponentials of our wave functions (keeping in mind that $\mathbf{r}$ is a position vector, so it depends on x and y). Here's the key thing to notice: Because the boundary conditions must hold everywhere on the boundary (where $z = 0$) at all times, the only way for Equation 4.22 to be true is if all the exponentials are equal and cancel out. Otherwise, a slight change in x or y would destroy the equality. The time factors are already equal always, which confirms that the transmitted, reflected, and incident frequencies must be equal, as we already reasoned out.

More interestingly, this equality of the exponentials also means that

$$\mathbf{k}_I \cdot \mathbf{r} = \mathbf{k}_R \cdot \mathbf{r} = \mathbf{k}_T \cdot \mathbf{r} \quad \text{at } z = 0. \tag{4.23}$$

If we actually expand those dot products out in terms of components, we could write this as

$$x\,(k_I)_x + y\,(k_I)_y = x\,(k_R)_x + y\,(k_R)_y = x\,(k_T)_x + y\,(k_T)_y \tag{4.24}$$

for all x and all y at $z = 0$. But we can say even more: Equation 4.24 can only possibly hold if each of the components are separately equal. This is easy to see if we just consider a particular point. If we set $x = 0$, then the equation becomes

$$(k_I)_y = (k_R)_y = (k_T)_y\,. \tag{4.25}$$

Similarly, setting $y = 0$ gives

$$(k_I)_x = (k_R)_x = (k_T)_x\,. \tag{4.26}$$

We can orient our coordinate system any way we'd like, so we can always make sure that $\mathbf{k}_I$ lies in the xz plane such that $(k_I)_y = 0$ (this is how it appears in Figure 4.2 if you assume that no components are pointing in or out of the page). Based on Equation 4.25, we can see that in that case, all three of the waves must lie in the xz plane. This is **one of the laws of geometric optics**: The incident, reflected, and transmitted wave vectors form a plane (called the **plane of incidence**), which also includes the normal to the surface (the z axis in our diagram).

Going back to Equation 4.26 and doing a little bit of trigonometry in Figure 4.2 implies that we can write

$$k_I \sin\theta_I = k_R \sin\theta_R = k_T \sin\theta_T, \tag{4.27}$$

where θ_I is the **angle of incidence**, θ_R is the **angle of reflection**, and θ_T is the **angle of refraction** (or sometimes the angle of transmission). Each of these angles is measured with respect to the normal of the boundary. We can use Equation 4.21 to eliminate the wave numbers, revealing two powerful laws. The angle of incidence is equal to the angle of relection,

$$\theta_I = \theta_R, \tag{4.28}$$

a fact known as the **law of reflection**. Additionally, the angles of incidence and refraction are related by **Snell's Law**,

$$n_1 \sin\theta_I = n_2 \sin\theta_T, \tag{4.29}$$

which is also sometimes known as the **law of refraction**.

Wow! We now have the three basic laws of geometric optics. I'm always kinda amazed by this derivation. Straight from Maxwell's Equations, we get basically all of geometric optics. Not only that, but we only used the most basic wave-like features of Maxwell's Equations, so these laws turn out to be even more general than just electromagnetic waves: other waves, such as sound waves, will follow analogous laws.

I guess now we can do some optics.

Chapter 5

The Basic Ideas of Geometric Optics

5.1 Rays and Reversibility in Geometric Optics

Now that we have the basic laws of geometrics optics, how will we apply them? While we've seen that the laws of geometric optics arise directly from the wave nature of light, geometric optics is so-called because for many purposes we can completely ignore that wave nature, instead focusing almost exclusively on geometry. In general, the study of optics is usually divided into two main categories:

- **Geometric optics**, also sometimes known as ray optics, refers to problems that can be tackled with only the laws of geometric optics and geometry. Practically speaking, this type of optical analysis is valid when diffraction effects (which we'll discuss later in the course) are negligible, which occurs when the light only encounters objects much bigger than its wavelength. Put another way, the basic study of lenses and mirrors is the usually purview of geometric optics.
- **Physical optics**, also sometimes known as wave optics, must be used if diffraction effects are relevant. In this case, we must fully consider all the wave-like behaviors of light, such as how waves will bend around edges and openings.

In geometric optics, we usually completely disregard the wave nature of light in favor of imagining that light is a perfect **ray**. A ray is simply the line along which the light travels; in terms of our wave derivation, it is the path along which the wave vector, $\mathbf{k}$, points. Rays are usually therefore simply drawn as arrows illustrating the path that light takes.

A careful inspection of the laws of geometric optics reveals another useful fact: any path that a light ray takes, even if that path contains reflections or refractions, is completely **reversible**.

5.2 Reflected and Refracted Images in Planes

To begin our practice using the laws of geometric optics, we will study a simple plane of material. The next several pages are an extract from Pedrotti, Pedrotti, & Pedrotti's [2] excellent discussion of the topic.

An extract from Chapter 2 of
Introduction to Optics, Third Edition
by Pedrotti, Pedrotti, & Pedrotti,
describing geometric optics applied to planes.

4 REFLECTION IN PLANE MIRRORS

Before discussing the formation of images in a general way, we discuss the simplest—and experientially, the most accessible—case of images formed by plane mirrors. In this context it is important to distinguish between specular *reflection* from a perfectly smooth surface and *diffuse reflection* from a granular or rough surface. In the former case, all rays of a parallel beam incident on the surface obey the law of reflection from a plane surface and therefore reflect as a parallel beam; in the latter case, though the law of reflection is obeyed locally for each ray, the microscopically granular surface results in

rays reflected in various directions and thus a diffuse scattering of the originally parallel rays of light. Every plane surface will produce some such scattering, since a perfectly smooth surface can only be approximated in practice. The treatment that follows assumes the case of specular reflection.

Consider the specular reflection of a single light ray OP from the xy-plane in Figure 7a. By the law of reflection, the reflected ray PQ remains within the plane of incidence, making equal angles with the normal at P. If the path OPQ is resolved into its x-, y-, and z-components, it is clear that the direction of ray OP is altered by the reflection only along the z-direction, and then in such a way that its z-component is simply reversed. If the direction of the incident ray is described by its unit vector, $\hat{\mathbf{r}}_1 = (x, y, z)$, then the reflection causes

$$\hat{\mathbf{r}}_1 = (x, y, z) \longrightarrow \hat{\mathbf{r}}_2 = (x, y, -z)$$

It follows that if a ray is incident from such a direction as to reflect sequentially from all three rectangular coordinate planes, as in the "corner reflector" of Figure 7b,

$$\hat{\mathbf{r}}_1 = (x, y, z) \longrightarrow \hat{\mathbf{r}}_2 = (-x, -y, -z)$$

and the ray returns precisely parallel to the line of its original approach. A network of such corner reflectors ensures the exact return of a beam of light—a headlight beam from highway reflectors, for example, or a laser beam from a mirror on the moon.

Image formation in a plane mirror is illustrated in Figure 8a. A point object S sends rays toward a plane mirror, which reflect as shown. The law of reflection ensures that pairs of triangles like SNP and $S'NP$ are equal, so all reflected rays appear to originate at the *image point* S', which lies along the normal line SN, and at such a depth that the *image distance* $S'N$ equals the *object distance* SN. The eye sees a point image at S' in exactly the same way it would see a real point object placed there. Since none of the actual rays of light lies below the mirror surface, the image is said to be a *virtual image*. The image S' *cannot* be projected on a screen as in the case of a *real image*. All points of an extended object, such as the arrow in Figure 8b, are imaged by a plane mirror in similar fashion: Each object point has its image point along its normal to the mirror surface and as far below the reflecting surface as the object point lies above the surface. Note that the image position does not depend on the position of the eye. Further, the construction of Figure 8b makes clear that the image size is identical with the object size, giving a *magnification* of unity. In addition, the transverse orientation of object and image are the same. A right-handed object, however, appears left-handed in its image. In Figure 8c, where the mirror does not lie directly below the object, the mirror plane may be extended to determine the position of the image as seen by an eye positioned to receive reflected rays originating at the object. Figure 8d illustrates multiple images of a point object O formed by two perpendicular mirrors. Images I_1 and I_2 result from single reflections in the two mirrors, but a third image I_3 results from sequential reflections from both mirrors.

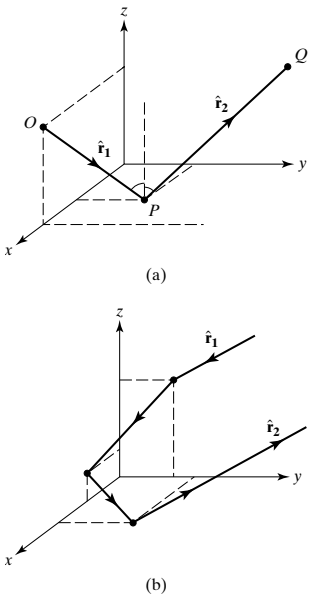

Figure 7 Geometry of a ray reflected from a plane.

5 REFRACTION THROUGH PLANE SURFACES

Consider light ray (1) in Figure 9a, incident at angle θ_1 at a plane interface separating two transparent media characterized, in order, by refractive indices n_1 and n_2. Let the angle of refraction be the angle θ_2. Snell's law, which now takes the form

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \tag{2}$$

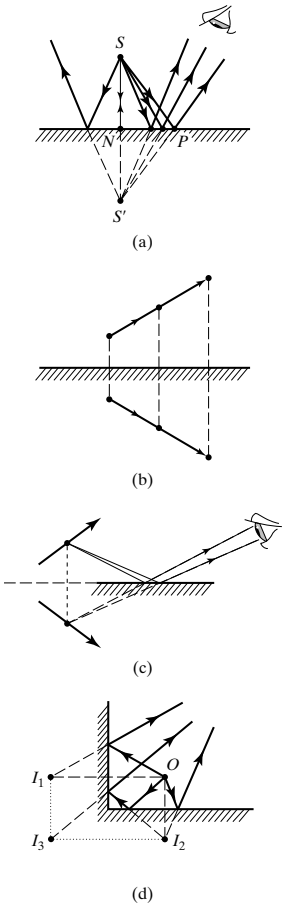

Figure 8 Image formation in a plane mirror.

requires an angle of refraction such that refracted rays bend away from the normal, as shown in Figure 9a, for rays 1 and 2, when $n_2 < n_1$. For $n_2 > n_1$, on the other hand, the refracted ray bends toward the normal. The law also requires that ray 3, incident normal to the surface ($\theta_1 = 0$), be transmitted without change of direction ($\theta_2 = 0$), regardless of the ratio of refractive indices.

In Figure 9a, the three rays shown originate at a source point S below an interface and emerge into an upper medium of lower refractive index, as in the case of light emerging from water ($n_1 = 1.33$) into air ($n_2 = 1.00$). A unique image point is not determined by these rays because they have no common intersection or virtual image point below the surface from which they appear to originate after refraction, as shown by the dashed line extensions of the refracted rays. For rays making a small angle with the normal to the surface, however, a reasonably good image can be located. In this approximation, where we allow only such *paraxial rays*[2] to form the image, the angles of incidence and refraction are both small, and the approximation

$$\sin\theta \cong \tan\theta \cong \theta \text{ (in radians)}$$

is valid. From Eq. (2), Snell's law can be approximated by

$$n_1 \tan\theta_1 \cong n_2 \tan\theta_2 \tag{3}$$

and taking the appropriate tangents from Figure 9b, we have

$$n_1\left(\frac{x}{s}\right) = n_2\left(\frac{x}{s'}\right)$$

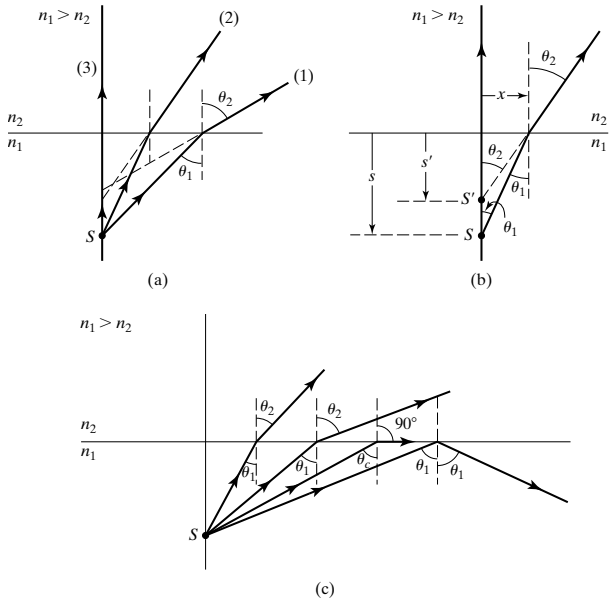

Figure 9 Geometry of rays refracted by a plane interface.

[2]In general, a paraxial ray is one that remains near the central axis of the image-forming optical system, thus making small angles with the optical axis.

The image point occurs at the vertical distance s' below the surface given by

$$s' = \left(\frac{n_2}{n_1}\right)s \tag{4}$$

where s is the corresponding depth of the object. Thus, objects underwater, viewed from directly overhead, appear to be nearer the surface than they actually are, since in this case $s' = (1/1.33)\, s = (3/4)\, s$. Even when the viewing angle θ_2 is not small, a reasonably good retinal image of an underwater object is formed because the aperture or pupil of the eye admits only a small bundle of rays while forming the image. Since these rays differ very little in direction, they will appear to originate from approximately the same image point. However, the depth of this image will not be 3/4 the object depth, as for paraxial rays, and in general will vary with the angle of viewing.

Rays from the object that make increasingly larger angles of incidence with the interface must, by Snell's law, refract at increasingly larger angles, as shown in Figure 9c. A critical angle of incidence θ_c is reached when the angle of refraction reaches 90°. Thus, from Snell's law,

$$\sin\theta_c = \left(\frac{n_2}{n_1}\right)\sin 90 = \frac{n_2}{n_1}$$

or

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \tag{5}$$

For angles of incidence $\theta_1 > \theta_c$, the incident ray experiences *total internal reflection*, as shown. For angle of incidence $\theta_1 < \theta_c$ both refraction and reflection occur. The reflected rays for this case are not shown in Figure 9c. This phenomenon is essential in the transmission of light along glass fibers by a series of total internal reflections. Note that the phenomenon does not occur unless $n_1 > n_2$, so that θ_c can be determined from Eq. (5).

We return to the nature of images formed by refraction at a plane surface when we deal with such refraction as a special case of refraction from a spherical surface.

Chapter 6

Basic Lenses and Mirrors: The Spherical Case and the Thin Lens Equation

6.1 More Complicated Optical Surfaces and Their Limitations

While optics like plane mirrors are useful for redirecting light and similar tasks, the broader question for the purposes of making astronomical observations is more difficult: How do we form an image, perhaps while applying magnification or using a much larger light collection surface than our eyes alone can acheive?

To begin our consideration of such task we will consider simple optical systems. When I refer to an optical system, I simply mean the combination of:

- a real object from which light emanates,
- a collection of optical elements through which that light propagates, and
- the image of the object that results at some other location as a consequence of the effects of the optical elements.

In essence, the optical elements will be able to create a real image if all rays that originate from one point on the object return to a single point after passing through the optical elements. To create an image of an extended object, this requirement must hold for every pair of object and image points.

While this requirement is easy to state, it is substantially more complicated to implement with real-world devices. Essentially no optical system meets this ideal, but many can approximate it, creating non-ideal images. We can group non-idealities into three broad categories: scattering, diffraction, and aberrations. All real-world optics will create some scattering and diffraction, because all optical elements have a finite size and are built from imprefectly smooth and homogeneous substances. We can thus never form a perfectly sharp image of an object. However, even if these two issues could be eliminated, aberrations would still prevent perfect images of any object being images in multiple wavelengths and with a finite spatial extent. Aberrations are fundmental distortions arising from the properties of the

objects themselves; generally, we cannot find a geometry that brings all possible light rays to a perfect focus.

However, so-called *Cartesian surfaces* are those mirror and lens shapes that form perfect images under certain restrictions (for example, a parabolic mirror will bring all on-axis parallel rays to a perfect focus). For the most part, the surfaces of interest are those that have cross sections that are conic sections: the sphere, ellipsoid, paraboloid, and hyperboloid. All of these optical element shapes can be roughly approximated at zeroth order as spherical surfaces. Further, most commerically-produced optical surfaces are spherical because of their ease of manufacture, and the spherical aberrations that result are accepted as a compromise when weighed against the relative ease of fabricating spherical surfaces. In the remainder of this chapte, therefore, we examine sphericalreflecting and refracting surfaces. Note that a plane surface, which we discussed previously, can be treated as a special case of a spherical surface in the limit that the radius of curvature R of the surface tends to infinity. In the next chapter, we will expand our discussion to paraboloids, which are one of the most common surfaces used in telescopes.

6.2 Spherical Optical Elements as Discussed by Pedrotti, Pedrotti, & Pedrotti

As in the last chapter, we will study this topic by reading an extract from Pedrotti, Pedrotti, & Pedrotti's [2] discussion of the topic. The following twelve pages are drawn from Chapter 2 of the third edition of their *Introduction to Optics*. (And if you don't like their writing, don't worry: they'll be gone in the next chapter.)

7 REFLECTION AT A SPHERICAL SURFACE

Spherical mirrors may be either concave or convex relative to an object point O, depending on whether the center of curvature C is on the same or opposite side of the reflecting surface. In Figure 15 the mirror shown is convex, and two

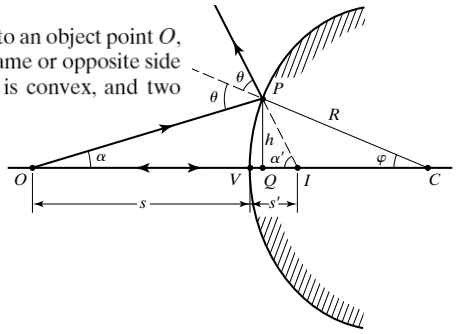

Figure 15 Reflection at a spherical surface.

rays of light originating at O are drawn, one normal to the spherical surface at its vertex V and the other an arbitrary ray incident at P. The first ray reflects back along itself; the second reflects at P as if from a plane tangent at P, satisfying the law of reflection. The two reflected rays diverge as they leave the mirror. The intersection of the two rays (extended backward) determines the image point I conjugate to O. The image is *virtual*, located behind the mirror surface.

Object and image distances from the vertex are shown as s and s', respectively. A perpendicular of height h is drawn from P to the axis at Q. We seek a relationship between s and s' that depends only on the radius of curvature R of the mirror. As we shall see, such a relation is possible only to first-order approximation of the sines and cosines of the angles made by the object and image rays to the spherical surface. This means that in place of the expansions

$$\sin\varphi = \varphi - \frac{\varphi^3}{3!} + \frac{\varphi^5}{5!} - \cdots$$

and

$$\cos\varphi = 1 - \frac{\varphi^2}{2!} + \frac{\varphi^4}{4!} + \cdots \tag{8}$$

we consider the first terms only and write

$$\sin\varphi \cong \varphi \quad \text{and} \quad \cos\varphi \cong 1 \tag{9}$$

relations that can be accurate enough if the angle φ is small enough.[4] This approximation leads to *first-order*, or *Gaussian*, optics, after Karl Friedrich Gauss, who in 1841 developed the foundations of the subject. Returning now to the problem at hand, notice that two angular relationships may be obtained from Figure 15, because the exterior angle of a triangle equals the sum of its interior angles. These are

$$\theta = \alpha + \varphi \quad \text{and} \quad 2\theta = \alpha + \alpha'$$

which combine to give

$$\alpha - \alpha' = -2\varphi \tag{10}$$

Using the small-angle approximation, the angles of Eq. (10) can be replaced by their tangents, yielding

$$\frac{h}{s} - \frac{h}{s'} = -2\frac{h}{R}$$

[4]For example, for angles φ around 10°, the approximation leads to errors around 1.5%.

where we have also neglected the axial distance VQ, small when angle φ is small. Cancellation of h produces the desired relationship,

$$\frac{1}{s} - \frac{1}{s'} = -\frac{2}{R} \tag{11}$$

If the spherical surface is chosen to be concave instead, the center of curvature would be to the left. For certain positions of the object point O, it is then possible to find a real image point also to the left of the mirror. In these cases, the resulting geometric relationship analogous to Eq. (11) consists of terms that are all positive. It is possible, by employing an appropriate sign convention, to represent all cases by the single equation

$$\frac{1}{s} + \frac{1}{s'} = -\frac{2}{R} \tag{12}$$

The sign convention to be used in conjunction with Eq. (12) is as follows. Assume the light propagates from left to right:

1. The *object distance s* is positive when O is to the left of V, corresponding to a real object. When O is to the right, corresponding to a virtual object, s is negative.
2. The *image distance s′* is positive when I is to the left of V, corresponding to a real image, and negative when I is to the right of V, corresponding to a virtual image.
3. The *radius of curvature R* is positive when C is to the right of V, corresponding to a convex mirror, and negative when C is to the left of V, corresponding to a concave mirror.

These rules[5] can be quickly summarized by noticing that positive object and image distances correspond to real objects and real images and that convex mirrors have positive radii of curvature. Applying Rule 2 to Figure 15, we see that the general Eq. (12) becomes identical with Eq. (11), a special case derived in conjunction with Figure 15. Virtual objects occur only with a sequence of two or more reflecting or refracting elements and are considered later.

The spherical mirror described by Eq. (12) yields, for a plane mirror with $R \to \infty$, $s' = -s$, as determined previously. The negative sign implies a virtual image for a real object. Notice also in Eq. (12) that object distance and image distance appear symmetrically, implying their interchangeability as conjugate points. For an object at infinity, incident rays are parallel and $s' = -R/2$, as illustrated in Figure 16a and b for both concave $(R < 0)$ and convex $(R > 0)$ mirrors. The image distance in each case is defined as the *focal length f* of the mirrors. Thus,

$$f = -\frac{R}{2}\begin{cases} >0, & \text{concave mirror} \\ <0, & \text{convex mirror} \end{cases} \tag{13}$$

and the mirror equation can be written, more compactly, as

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{14}$$

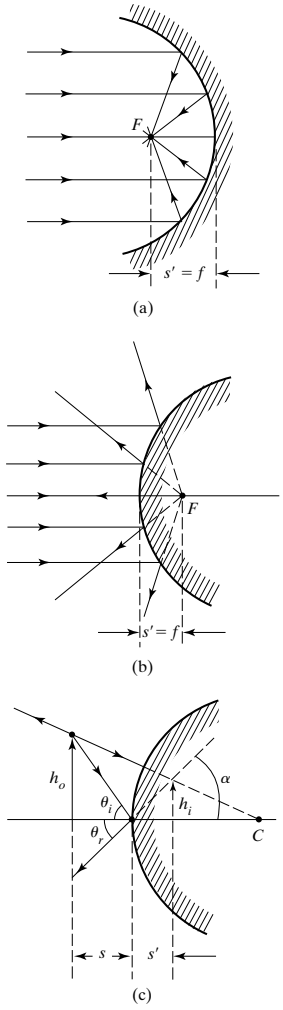

Figure 16 Location of focal points (a) and (b) and construction to determine magnification (c) of a spherical mirror.

[5]Although this set of sign conventions is widely used, the student is cautioned that other schemes exist. No one with a continuing involvement in optics can hope to escape confronting other conventions, nor should the matter be beyond the mental flexibility of the serious student to accommodate.

The focal point F, located a focal length f from the vertex of the mirror, and shown in Figure 16a and b, serves as an important construction point in graphical ray-tracing techniques, which we discuss following Example 1.

In Figure 16c, a construction is shown that allows the determination of the transverse magnification. The object is an extended object of transverse dimension h_o. The image of the top of the object arrow is located by two rays whose behavior on reflection is known. The ray incident at the vertex must reflect to make equal angles with the axis. The other ray is directed toward the center of curvature along a normal and so must reflect back along itself. The intersection of the two reflected rays occurs behind the mirror and locates a virtual image of dimension h_i there. Because of the equality of the three angles shown, it follows that

$$\frac{h_o}{s} = \frac{h_i}{s'}$$

The lateral magnification m is defined by the ratio of lateral image size to corresponding lateral object size, so that

$$|m| = \frac{h_i}{h_o} = \frac{s'}{s} \tag{15}$$

Extending the sign convention to include magnification, we assign a $(+)$ magnification to the case where the image has the same orientation as the object and a $(-)$ magnification where the image is inverted relative to the object. To produce a $(+)$ magnification in the construction of Figure 16c, where s' must itself be negative, we modify Eq. (15) to give the general form

$$m = -\frac{s'}{s} \tag{16}$$

The following example illustrates the correct use of the sign convention.

Example 1

An object 3 cm high is placed 20 cm from (a) a convex and (b) a concave spherical mirror, each of 10-cm focal length. Determine the position and nature of the image in each case.

Solution

a. Convex mirror: $f = -10$ cm and $s = +20$ cm.

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \quad \text{or} \quad s' = \frac{fs}{s - f} = \frac{(-10)(20)}{(20) - (-10)} = -6.67 \text{ cm}$$

$$m = -\frac{s'}{s} = -\frac{-6.67}{20} = +0.333 = \frac{1}{3}$$

The image is virtual (because s' is negative), 6.67 cm to the right of the mirror vertex, and is erect (because m is positive) and $\frac{1}{3}$ the size of the object, or 1 cm high.

b. Concave mirror: $f = +10$ cm and $s = +20$ cm.

$$s' = \frac{fs}{s - f} = \frac{(10)(20)}{20 - 10} = +20 \text{ cm}$$

$$m = -\frac{s'}{s} = -\frac{20}{20} = -1$$

The image is real (because s' is positive), 20 cm to the left of the mirror vertex, and is inverted (because m is negative) and the same size as the object, or 3 cm high. Image and object happen to be at $2f = 20$ cm, the center of curvature of the mirror.

The location and nature of the image formed by a mirror can be determined by graphical ray-trace techniques. Figure 17 illustrates how three key rays—labeled 1, 2, and 3—each leaving a point P at the tip of an object, can be drawn to locate the conjugate image point P'. In fact, under the conditions for which Eqs. (12) through (16) are valid, the paths of *any* two rays leaving P are sufficient to locate the conjugate image point P'. A third ray serves as a convenient check on the accuracy of the first two chosen rays. The three key rays discussed in connection with Figure 17 are chosen as the basis of the graphical ray-trace technique because, once the mirror center of curvature C, the focal point F, and vertex V are located along the optical axis of a spherical mirror, these three rays can be drawn using only a straightedge device. The conjugate image point P' marks the tip of the image—the entire image then lies between P' and the point on the optical axis directly above or below P'.

Refer to Figure 17a, b, and c in connection with the following description of how the three key rays can be drawn. Note the difference in each ray

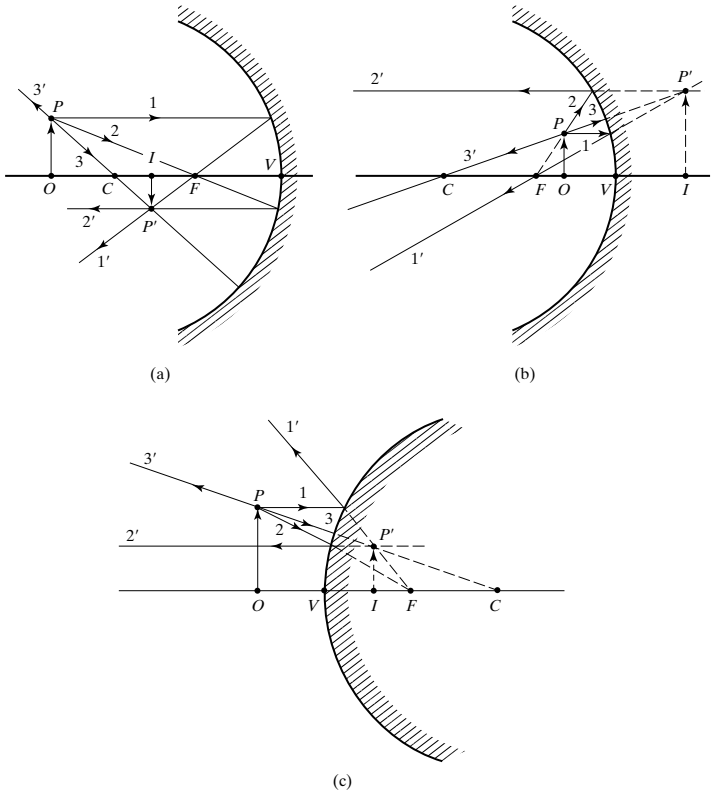

Figure 17 Ray diagrams for spherical mirrors. (a) Real image, concave mirror. The object distance is greater than the focal length. (b) Virtual image, concave mirror. The object distance is less than the focal length. (c) Virtual image, convex mirror.

trace, depending on the object location before or after points C and F, and on the geometry of the mirror surface, concave or convex.

- *Ray 1*. This ray leaves point P as a ray parallel to the optical axis, strikes the mirror, reflects and passes through the focal point F of a *concave* mirror—as in Figure 17a and b. Or, as in Figure 17c, it strikes a *convex* mirror and reflects as if it came from the focal point F behind the mirror. In each case, after reflection this ray is labeled $1'$.
- *Ray 2*. This ray leaves point P, passes through F, strikes a *concave* mirror, and is reflected as a ray parallel to the optical axis, as in Figure 17a. Or, as in Figure 17b, it leaves point P as if it is coming from the point F to its left (dotted line), strikes the *concave* mirror, and reflects as a parallel ray. Or, as in Figure 17c, for a *convex* mirror, the ray leaves point P heading toward focal point F behind the mirror, strikes the mirror, and reflects as a parallel ray. In each case, after reflection, this ray is labeled $2'$.
- *Ray 3*. This ray leaves point P in Figure 17a, passes through point C for the *concave* mirror, strikes the mirror, and reflects back along itself. Or, as in Figure 17b—still for a *concave* mirror—ray 3 appears to come from the point C to its left, strikes the mirror, and reflects back along itself. Or, as in Figure 17c, for a *convex* mirror, it heads toward point C behind the mirror, strikes the mirror, and reflects back along itself. In each case, after reflection, this ray is labeled $3'$.

To understand how these rays locate the conjugate image point P' that marks the tip of the image, it is useful to imagine that these three rays arrive at the eye of one viewing the image. For the case shown in Figure 17a, the three rays $1'$, $2'$, and $3'$ intersect at a real image point as they progress away from the mirror and toward the viewer. For the arrangements shown in Figure 17b and 17c, the rays $1'$, $2'$, and $3'$ appear to originate from a point of intersection (a virtual image point) located behind the mirror. The real or apparent point of intersection is interpreted as the emanation point of these rays. That is, the viewer "sees" the tip of an image at point P'.

8 REFRACTION AT A SPHERICAL SURFACE

We turn now to a similar treatment of refraction at a spherical surface, choosing in this case the concave surface of Figure 18. Two rays are shown emanating from object point O. One is an axial ray, normal to the surface at its vertex and so refracted without change in direction. The other ray is an arbitrary ray incident at P and refracting there according to Snell's law,

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \tag{17}$$

The two refracted rays appear to emerge from their common intersection, the image point I. In triangle CPO, the exterior angle $\alpha = \theta_1 + \varphi$. In triangle CPI, the exterior angle $\alpha' = \theta_2 + \varphi$. Approximating for paraxial rays and substituting for θ_1 and θ_2 in Eq. (17), we have

$$n_1(\alpha - \varphi) = n_2(\alpha' - \varphi) \tag{18}$$

Next, writing the tangents for the angles by inspection of Figure 18, where again we may neglect the distance QV in the small angle approximation,

$$n_1\left(\frac{h}{s} - \frac{h}{R}\right) = n_2\left(\frac{h}{s'} - \frac{h}{R}\right)$$

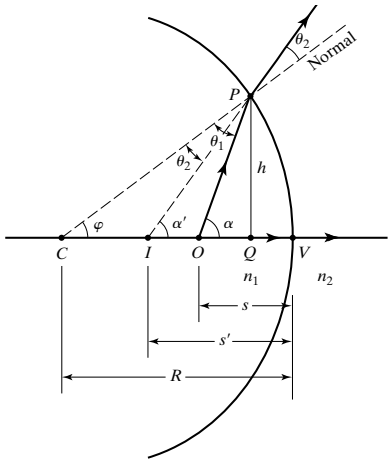

Figure 18 Refraction at a spherical surface for which $n_2 > n_1$.

or

$$\frac{n_1}{s} - \frac{n_2}{s'} = \frac{n_1 - n_2}{R} \tag{19}$$

Employing the *same sign convention* as introduced for mirrors (i.e., positive distances for real objects and images and negative distances for virtual objects and images), the virtual image distance $s' < 0$ and the radius of curvature $R < 0$. If these negative signs are understood to apply to these quantities for the case of Figure 18, a general form of the refraction equation may be written as

$$\frac{n_1}{s} + \frac{n_2}{s'} = \frac{n_2 - n_1}{R} \tag{20}$$

which holds equally well for convex surfaces. When $R \to \infty$, the spherical surface becomes a plane refracting surface, and

$$s' = -\left(\frac{n_2}{n_1}\right)s \tag{21}$$

where s' is the apparent depth determined previously. For a real object ($s > 0$), the negative sign in Eq. (21) indicates that the image is virtual. The lateral magnification of an extended object is simply determined by inspection of Figure 19. Snell's law requires, for the ray incident at the vertex V and in the small-angle approximation, $n_1\theta_1 = n_2\theta_2$ or, using tangents for angles,

$$n_1\left(\frac{h_o}{s}\right) = n_2\left(\frac{h_i}{s'}\right)$$

The lateral magnification is, then,

$$m = \frac{h_i}{h_o} = -\frac{n_1 s'}{n_2 s} \tag{22}$$

where the negative sign is attached to give a negative value corresponding to an inverted image. For the case of a plane refracting surface, Eq. (21) may

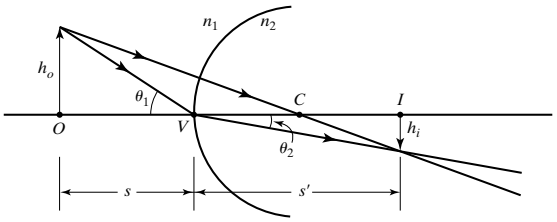

Figure 19 Construction to determine lateral magnification at a spherical refracting surface.

be incorporated into Eq. (22), giving $m = +1$. Thus, the images formed by plane refracting surfaces have the same lateral dimensions and orientation as the object.

Example 2

As an extended example of refraction by spherical surfaces, refer to Figure 20. In (a), a real object is positioned in air, 30 cm from a convex spherical surface of radius 5 cm. To the right of the interface, the refractive index is that of water. Before constructing representative rays, we first find the image distance and lateral magnification of the image, using Eqs. (20) and (22). Equation (20) becomes

$$\frac{1}{30} + \frac{1.33}{s'_1} = \frac{1.33 - 1}{5}$$

giving $s'_1 = +40$ cm. The positive sign indicates that the image is real and so is located to the right of the surface, where real rays of light are refracted. Equation (22) becomes

$$m = -\frac{(1)(+40)}{(1.33)(+30)} = -1$$

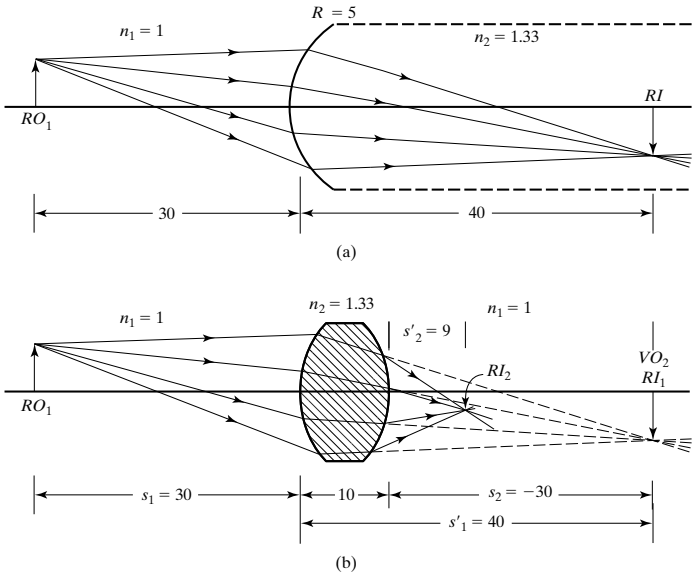

Figure 20 Example of refraction by spherical surfaces. (All distances are in cm.) (a) Refraction by a single spherical surface. (b) Refraction by a thick lens. Subscripts 1 and 2 refer to refractions at the first and second surfaces, respectively.

indicating an inverted image, equal in size to that of the object. Figure 20a shows the image, as well as several rays, which are now determined. In this example we have assumed that the medium to the right of the spherical surface extends far enough so that the image is formed inside it, without further refraction. Let us suppose now (Figure 20b) that the second medium is only 10 cm thick, forming a *thick lens*, with a second, concave spherical surface, also of radius 5 cm. The refraction by the first surface is, of course, unaffected by this change. Inside the lens, therefore, rays are directed as before to form an image 40 cm to the right of the first surface. However, these rays are intercepted and refracted by the second surface to produce a different image, as shown. Since the convergence of the rays striking the second surface is determined by the position of the first image, its location now specifies the appropriate object distance to be used for the second refraction. We call the real image formed by surface (1) a *virtual object* for surface (2). Then, by the *sign convention established previously*, we make the virtual object distance, relative to the second surface, a negative quantity when using Eqs. (20) and (22). For the second refraction, then, Eq. (20) becomes

$$\frac{1.33}{-30} + \frac{1}{s'_2} = \frac{1 - 1.33}{-5}$$

or $s' = +9$ cm. The magnification, according to Eq. (22), is

$$m = \frac{(-1.33)(+9)}{(1)(-30)} = +\frac{2}{5}$$

The final image is, then, 2/5 the lateral size of its (virtual) object and apears with the same orientation. Relative to the original object, the final image is 2/5 as large and inverted.

In general, whenever a train of reflecting or refracting surfaces is involved in the processing of a final image, the individual reflections and/or refractions are considered in the order in which light is actually incident upon them. The object distance of the nth step is determined by the image distance for the $(n - 1)$th step. If the image of the $(n - 1)$ step is not actually formed, it serves as a *virtual object* for the nth step.

9 THIN LENSES

We now apply the preceding method to discover the thin-lens equation. As in the example of Figure 20, two refractions at spherical surfaces are involved. The simplification we make is to neglect the thickness of the lens in comparison with the object and image distances, an approximation that is justified in most practical situations. At the first refracting surface, of radius R_1,

$$\frac{n_1}{s_1} + \frac{n_2}{s'_1} = \frac{n_2 - n_1}{R_1} \tag{23}$$

and at the second surface, of radius R_2,

$$\frac{n_2}{s_2} + \frac{n_1}{s'_2} = \frac{n_1 - n_2}{R_2} \tag{24}$$

We have assumed that the lens faces the same medium of refractive index n_1 on both sides. Now the second object distance, in general, is given by

$$s_2 = t - s'_1 \tag{25}$$

where t is the thickness of the lens. Notice that this relationship produces the correct sign of s_2, as in Figure 20, and also when the intermediate image falls inside or to the left of the lens. In the thin-lens approximation, neglecting t,

$$s_2 = -s_1' \tag{26}$$

When this value of s_2 is substituted into Eq. (24) and Eqs. (23) and (24) are added, the terms n_2/s_1' cancel and there results

$$\frac{n_1}{s_1} + \frac{n_1}{s_2'} = (n_2 - n_1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

Now s_1 is the original object distance and s_2' is the final image distance, so we may drop their subscripts and write simply

$$\frac{1}{s} + \frac{1}{s'} = \frac{n_2 - n_1}{n_1}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{27}$$

The *focal length* of the thin lens is defined as the image distance for an object at infinity, or the object distance for an image at infinity, giving

$$\frac{1}{f} = \frac{n_2 - n_1}{n_1}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{28}$$

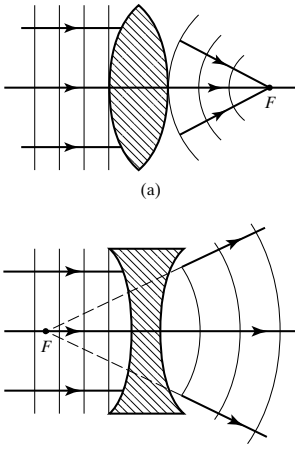

Figure 21 Lens action on plane wavefronts of light. (a) Converging lens (positive focal length). (b) Diverging lens (negative focal length).

Equation (28) is called the *lensmaker's equation* because it predicts the focal length of a lens fabricated with a given refractive index and radii of curvature and used in a medium of refractive index n_1. In most cases, the ambient medium is air, and $n_1 = 1$. The thin-lens equation, in terms of the focal length, is then

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{29}$$

Wavefront analysis for plane wavefronts, as shown in Figure 21, indicates that a lens thicker in the middle causes convergence, and one thinner in the middle causes divergence of the incident parallel rays. The portion of the wavefront that must pass through the thicker region is delayed relative to the other portions. Converging lenses are characterized by positive focal lengths and diverging lenses by negative focal lengths, as is evident from the figure, where the images are real and virtual, respectively.

Sample ray diagrams for converging (or *convex*) and diverging (or *concave*) lenses are shown in Figure 22. The thin lenses are best represented, for purposes of ray construction, by a vertical line with edges suggesting the general shape of the lens—ordinary arrowheads for converging lenses, inverted arrowheads for diverging lenses. Graphical methods of locating images, as with spherical mirrors in Figure 17, make use of *three key rays*. This procedure is outlined next and illustrated in Figures 22 and 23. The three rays leaving the tip of the object change direction due to refraction at the thin-lens interfaces. The redirected rays can be used to locate the image.

- *Ray 1*. A ray leaving the tip of the object, parallel to the optical axis, undergoing refraction at the lens surfaces and passing through the *right* focal point F of a *converging* lens, as in Figure 22a. Or, as in Figure 22b, a parallel ray which refracts at the lens surfaces as if coming directly from the *left* focal point F of a *diverging* lens.
- *Ray 2*. A ray leaving the tip of the object and passing through the *left* focal point F of a *converging* lens, undergoing refraction at the lens surfaces,

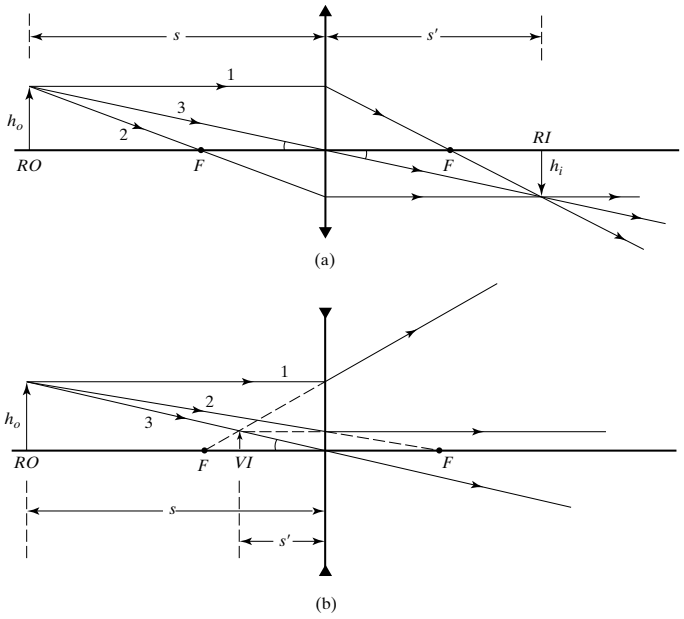

Figure 22 Ray diagrams for image formation by a convex lens (a) and a concave lens (b).

and emerging parallel to the axis as in Figure 22a. Or, as in Figure 22b, a ray leaving the tip of the object, directed toward the *right* focal point F of a *diverging* lens, undergoing refraction at the lens and emerging parallel to the axis.

- *Ray 3*. A ray leaving the tip of the object and passing directly through the center of a converging or diverging lens, emerging unaltered, as in Figure 22a or 22b.

The viewer, located at the far right in Figure 22a and 22b, receives these rays as if they have come directly from an object and so "sees" the tip of the image at the point where the backwards extensions of these rays either intersect or appear to intersect. Any two rays are sufficient to locate the image; the third ray may be drawn as a check on the accuracy of the graphical trace.

In constructing ray diagrams, as in Figure 22, observe that, except for the central ray (ray 3), each ray refracted by a convex lens bends toward the axis and each ray refracted by a concave lens bends away from the axis. From either diagram, the angles subtended by object and image at the center of the lens are seen to be equal. For either the real image *RI* in (a) or the virtual image *VI* in (b), it follows that

$$\frac{h_o}{s} = \frac{h_i}{s'}$$

and lateral magnification

$$|m| = \left|\frac{h_i}{h_o}\right| = \left|\frac{s'}{s}\right|$$

In accordance with the sign convention adopted here, the magnification should be the negative of the ratio of the image and object distances since, in case (a), $s > 0$, $s' > 0$, and $m < 0$ because the image is inverted; in case (b),

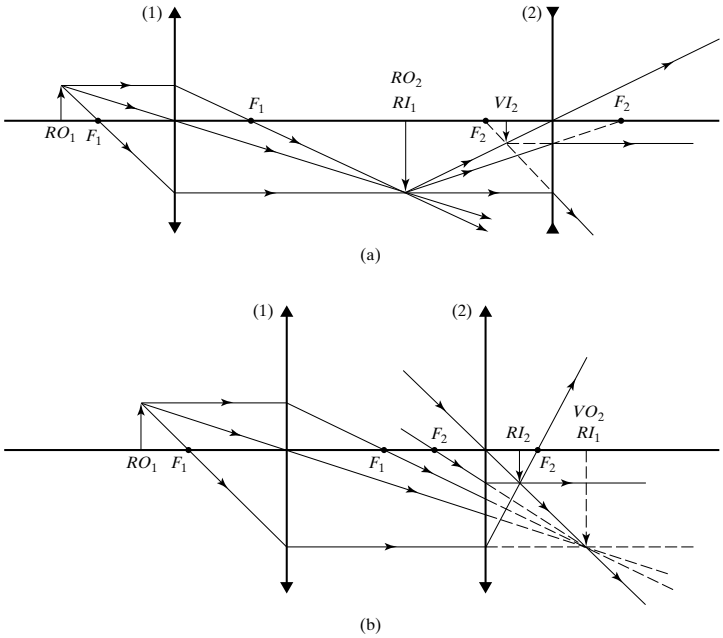

Figure 23 (a) Formation of a virtual image VI_2 by a two-element train of a convex lens (1) and concave lens (2). (b) Formation of a real image RI_2 by a train of two convex lenses. The intermediate image RI_1 serves as a virtual object VO_2 for the second lens.

$s > 0$, $s' < 0$, and $m > 0$. In either case, then,

$$m = -\frac{s'}{s} \tag{30}$$

Further ray-diagram examples for *a train of two lenses* are illustrated in Figure 23 and a calculation involving image formation in two lenses is given in Example 3.

Example 3

Find and describe the intermediate and final images produced by a two-lens system such as the one sketched in Figure 23a. Let $f_1 = 15$ cm, $f_2 = 15$ cm, and their separation be 60 cm. Let the object be 25 cm from the first lens, as shown.

Solution

The first lens is convex: $f_1 = +15$ cm, $s_1 = 25$ cm.

$$\frac{1}{s_1} + \frac{1}{s_1'} = \frac{1}{f} \quad \text{or} \quad s_1' = \frac{s_1 f}{s_1 - f} = \frac{(25)(15)}{25 - 15} = +37.5 \text{ cm}$$

$$m_1 = -\frac{s_1'}{s_1} = -\frac{37.5}{25} = -1.5$$

Thus, the first image is real (because s_1' is positive), 37.5 cm to the right of the first lens, inverted (because m is negative), and 1.5 times the size of the object.

The second lens is concave: $f_2 = -15$ cm. Since real rays of light diverge from the first real image, it serves as a real object for the second lens, with $s_2 = 60 - 37.5 = +22.5$ cm to the left of the lens. Then,

$$s_2' = \frac{s_2 f}{s_2 - f} = \frac{(22.5)(-15)}{(22.5) - (-15)} = -9 \text{ cm}$$

$$m_2 = -\frac{s_2'}{s_2} = -\frac{-9}{22.5} = +0.4$$

Thus, the final image is virtual (because s_2' is negative), 9 cm to the *left* of the seconds lens, erect *with respect to its own object* (because m is positive), and 0.4 times its size. The *overall* magnification is given by $m = m_1 m_2 = (-1.5)(0.4) = -0.6$. Thus, the final image is inverted relative to the *original* object and 6/10 its lateral size. All these features are exhibited qualitatively in the ray diagram of Figure 23a.

Table 1 and Figure 24 provide a convenient summary of image formation in lenses and mirrors.

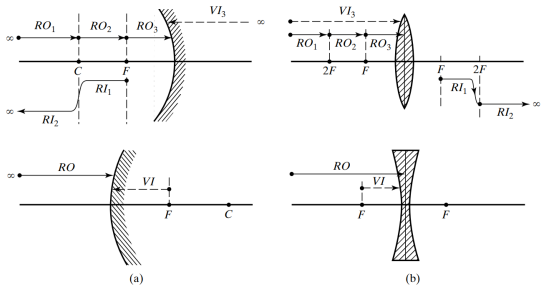

Figure 24 Summary of image formation by (a) spherical mirrors and (b) thin lenses. The location, nature, magnification, and orientation of the image are indicated or suggested. The letters *R* and *V* refer to *real* and *virtual*, *O* and *I* to *object* and *image*. Changes in elevation of the horizontal lines suggest the magnification in the various regions.

TABLE 1 SUMMARY OF GAUSSIAN MIRROR AND LENS FORMULAS

	Spherical surface	Plane surface
Reflection	$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}, f = -\frac{R}{2}$ $m = -\frac{s'}{s}$ Concave: $f > 0, R < 0$ Convex : $f < 0, R > 0$	$s' = -s$ $m = +1$
Refraction Single surface	$\frac{n_1}{s} + \frac{n_2}{s'} = \frac{n_2 - n_1}{R}$ $m = -\frac{n_1 s'}{n_2 s}$ Concave: $R < 0$ Convex : $R > 0$	$s' = -\frac{n_2}{n_1} s$ $m = +1$
Refraction Thin lens	$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$ $\frac{1}{f} = \frac{n_2 - n_1}{n_1}\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$ $m = -\frac{s'}{s}$ Concave: $f < 0$ Convex : $f > 0$	

References

[1] David J. Griffiths. *Introduction to Electrodynamics*. 2017.

[2] F.L. Pedrotti, L.M. Pedrotti, and L.S. Pedrotti. *Introduction to Optics*. Cambridge University Press, 2017.