Advanced Theoretical Physics

A Historical Perspective

Nick Lucid

June 2015

Last Updated: January 2023

Contents

Preface

In November of 2009, a friend asked me about Lagrangian mechanics and about a week later I returned to him having written Sections 4.2 and 4.3. The writing made sense to him and it occurred to me that I enjoyed the experience. It was relieving to get the knowledge out of my head and it felt rewarding to pass it onto someone else. In fact, I was so consumed by it, that I continued writing until I had written all of Chapter 4. It was at that moment that I decided to write this book.

I realized writing Chapter 4 there were many things I hadn't learned in my undergraduate physics courses but that my graduate professors expected me to already know. This made graduate school particularly challenging. It happens for a lot of reasons. Sometimes the teacher decides to focus on other material or runs out of time. Sometimes I was simply too busy taking several physics classes at once to worry about certain details. Other times the teacher assigns it as a reading assignment and, let's be honest, how many students actually do assigned reading? Even if you do the reading, sometimes the author dances around it like they either don't understand it themselves or they think it'll be fun for you to figure it out on your own. No matter what, we don't learn it when we're supposed to and it would be helpful if there was some person or some book that just says it plainly.

This book is intended to be just that. I wrote it primarily for advanced readers. You might want to read this book if:

- you're an undergraduate physics student planning on attending graduate school,
- you're a graduate physics student but feel like you're missing something, or
- you're someone who likes a challenge.

The point being, this book is not intended for anyone without at least *some* background in basic calculus and introductory physics.

The chapters of this book correspond to major topics, so some of them can get rather long. If a particular physics topic requires *a lot* of mathematical background, then development of that math will be in its own chapter preceding the physics topic. For example, vector calculus (Chapter 3) precedes electrodynamics (Chapter 5) and tensor analysis (Chapter 6) precedes relativity (Chapters 7 and 8). The topics are also in a somewhat historical order and include a bit of historical information to put them in context with each other. Historical context can give you a deeper insight into a topic and understanding how long it took the scientific community to develop something can make you feel a little better about maybe not understanding it immediately.

With the exception of Chapter 1, all chapters contain worked examples where helpful. Some of those examples also make use of numerical methods which can be found in Appendix A. Reading textbooks and other trade books on these topics, I often get frustrated by how many steps are missing from examples and derivations. As you read this book, you'll find that I make a point to include as many steps as possible and clearly explain any steps I don't show mathematically. Also, with so many different topics in one place, there are times where I avoid traditional notation in favor keeping a consistent notation throughout the book. Frankly, some traditional choices for symbols are terrible anyway.

Acknowledgments

I'd like to acknowledge Nicholas Arnold for proofreading this book and Jesse Mason for asking me that question about Lagrangian mechanics all those years ago.

Chapter 1

Coordinate Systems

Coordinate systems are something we get used to using very early on in mathematics. Their existence, among other things, is drilled into us with unyielding resolve. This can have undesired consequences such as preconception, so before we get into the thick of our discussion I'd like to make a few things clear.

- *Math is not the language of the universe.* As much as some of us like to think we're speaking the universe's language when we apply math to it, this simply isn't the case. The universe does what it does without concern for number crunching of any kind. It doesn't add, subtract, multiply, or divide. It doesn't take derivatives or integrals. As we see throughout this book, there are plenty of cases in which an exact solution is not attainable. What we do see is the universe is a relatively ordered place and mathematics is the most ordered tool we possess, so they seem to correlate.

- *The universe doesn't give preference to any particular coordinate system.* Coordinate systems are a tool of mathematics, which we've already seen the universe doesn't concern itself with. We can choose any coordinate system we wish for a given scenario. However, mathematical problems are more difficult to solve (or sometimes unsolvable) in a particular coordinate system. There is usually a best choice given the details of the scenario that will maximize the ease at which we can solve it, but this does not imply the universe had anything to do with the choice we're making.

Figure 1.1: René Descartes

- *When working with the specific, we always need to concern ourselves with a coordinate system.* This is why the importance of a coordinate system is stressed throughout our educational careers. We can't apply math at all without, at the very least, a point of reference (e.g. zero, infinity, initial conditions, boundary conditions, etc.). There may be a time when our math is so general it becomes coordinate system independent, which is good since the universe doesn't have one anyway. However, whenever we apply that work to something specific, the coordinates will always come into play.

With all this in mind, we have quite a few options. I've given some of the basic ones in the following sections.

1.1 Cartesian

The Cartesian coordinate system was developed by René Descartes (Latin: Renatus Cartesius). He published the concept in his work *La Géométrie* in 1637. The idea of uniting algebra with geometry as Descartes had resulted in drastic positive consequences on the development of mathematics, particularly the soon to be invented calculus.

This system of coordinates is the most basic we have consisting of, in general, three numbers to represent location: x, y, and z. It is a form of **rectilinear** coordinates, which is simply a grid of straight lines. We can represent this position as a position vector,

$$\vec{r} \equiv x\hat{x} + y\hat{y} + z\hat{z}\,, \tag{1.1.1}$$

where $\hat{x}$, $\hat{y}$, and $\hat{z}$ represent the directions along each of the axes. This

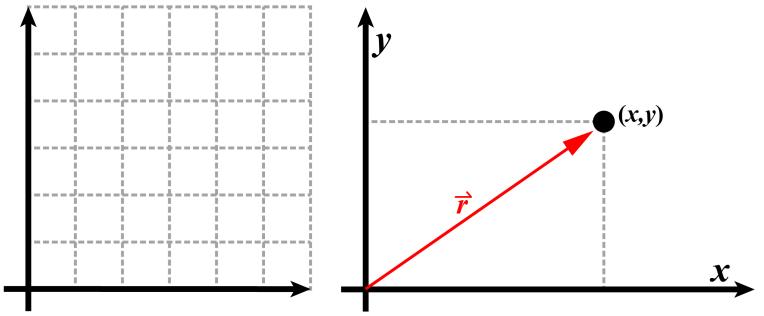


Figure 1.2: This is the Cartesian plane (i.e. the xy-plane or $\mathbb{R}^2$). The left shows the grid living up to it's rectilinear name. The right shows an arbitrary position vector in this system.

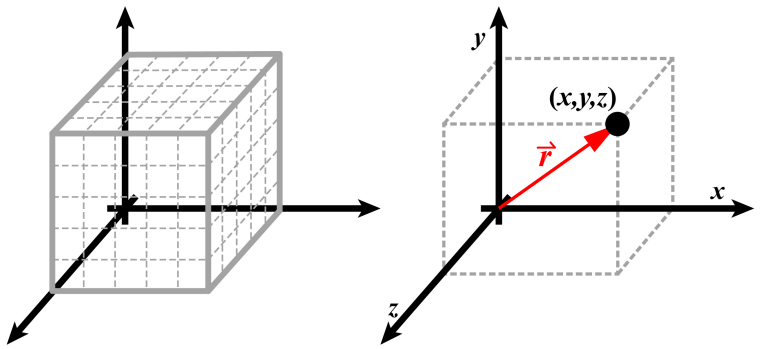


Figure 1.3: This is Cartesian space (i.e. $\mathbb{R}^3$). The left shows the grid living up to it's rectilinear name. The right shows an arbitrary position vector in this system.

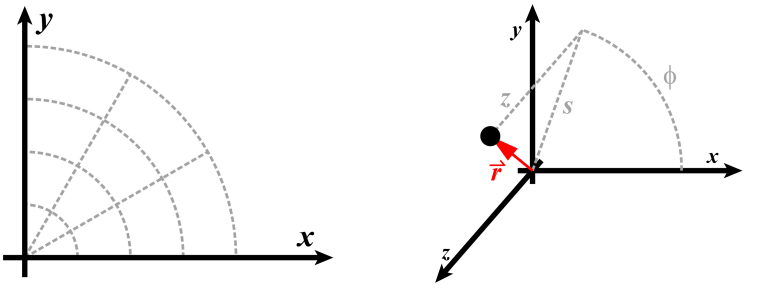


Figure 1.4: This is cylindrical coordinates. The left shows the curvilinear grid in the xy-plane (i.e. polar coordinates). The right shows an arbitrary position vector in this system where the coordinates are also labeled.

coordinate option might make the most sense, but it doesn't always make problem solving easier. For those cases, we have some more specialized options.

1.2 Polar and Cylindrical

Polar coordinates are a form of **curvilinear** coordinates, which is simply a grid where one or more of the lines are curved. In polar, we take a straight line from the origin to a point and refer to that distance as s (or sometimes r). Orientation of this line is determined by an angle, ϕ (or sometimes θ), measured from the Cartesian x-axis. Polar is expanded to cylindrical coordinates by adding an extra z value for three dimensions as shown in Figure 1.4. The terms **polar** and **cylindrical** are often used interchangeably.

As the previous paragraph suggests, there is a way to transform back and forth between cylindrical and Cartesian coordinates. The transformation equations are

$$\left\{\begin{array}{rcl} x &=& s\cos\phi \\ y &=& s\sin\phi \\ z &=& z \end{array}\right\} \tag{1.2.1}$$

or in reverse

$$\left\{\begin{array}{rcl} s &=& \sqrt{x^2+y^2} \\ \phi &=& \arctan\left(\frac{y}{x}\right) \\ z &=& z \end{array}\right\}. \tag{1.2.2}$$

We can also use Eq. 1.1.1, to find the corresponding unit vectors. Since we know $\hat{s} = \vec{s}/s$ and $\hat{\phi}$ must be perpendicular to $\hat{s}$ (initially with a positive y-component), the cylindrical unit vectors can be written as

$$\left.\begin{array}{rcl} \hat{s} &=& \cos\phi\,\hat{x} + \sin\phi\,\hat{y} \\ \hat{\phi} &=& -\sin\phi\,\hat{x} + \cos\phi\,\hat{y} \\ \hat{z} &=& \hat{z} \end{array}\right\}. \tag{1.2.3}$$

Writing these in matrix form, we have

$$\begin{bmatrix} \hat{s} \\ \hat{\phi} \\ \hat{z} \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix}. \tag{1.2.4}$$

Based on Eq. 1.2.4, we can see that all it takes to find the Cartesian unit vectors in terms of the cylindrical ones is to multiply through by the inverse of the coefficient matrix. This results in

$$\begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \hat{s} \\ \hat{\phi} \\ \hat{z} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} \tag{1.2.5}$$

$$\begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{s} \\ \hat{\phi} \\ \hat{z} \end{bmatrix} = \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix}. \tag{1.2.6}$$

Therefore, in equation form, they are

$$\left.\begin{array}{rcl} \hat{x} &=& \cos\phi\,\hat{s} - \sin\phi\,\hat{\phi} \\ \hat{y} &=& \sin\phi\,\hat{s} + \cos\phi\,\hat{\phi} \\ \hat{z} &=& \hat{z} \end{array}\right\}. \tag{1.2.7}$$

This set of coordinates is particularly useful when dealing with cylindrical symmetry (e.g. rotating rigid bodies, strings of mass, lines of charge, long straight currents, etc.)

1.3 Spherical

Just as with polar, spherical coordinates are a form of **curvilinear** coordinates. However, rather than having two straight lines and one curved, it's

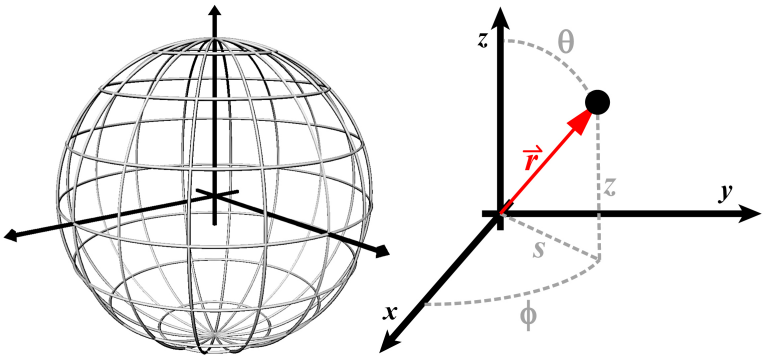


Figure 1.5: This is spherical coordinates. The left shows the grid you'd find on a surface of constant r and the right shows the arbitrary position vector. The orientation of the Cartesian system is different than it was in Figure 1.4, so the cylindrical coordinates are also shown for clarity.

the other way around. Position in spherical coordinates is determined by a radial distance, r (or sometimes ρ), from the origin and two independent angles: θ and ϕ. The definition of these two angles varies by application and field of study, but usually in physics (particularly in the Chapter 5 with electrodynamics) we define them as follows.

The angle ϕ is from the positive x-axis around in the xy-plane spanning values from 0 to 2π and θ is from the positive z-axis around in the sz-plane (at least that's what we'd call it in cylindrical coordinates) spanning values from 0 to π as show in Figure 1.5. We also see from Figure 1.5 the grid showing both angles is very similar to the latitude and longitude grid we've placed on the Earth.

The coordinate transformations from spherical to Cartesian coordinates are given by

$$\left\{\begin{array}{rcl} x &=& r\sin\theta\cos\phi \\ y &=& r\sin\theta\sin\phi \\ z &=& r\cos\theta \end{array}\right\} \tag{1.3.1}$$

where we have taken the cylindrical coordinates and made the substitutions of

$$\left\{\begin{array}{rcl} s &=& r\sin\theta \\ \phi &=& \phi \\ z &=& r\cos\theta \end{array}\right\}, \tag{1.3.2}$$

which transform from spherical to cylindrical. The origin of these substitutions can be easily seen in Figure 1.5 where we have included the cylindrical coordinates for clarity. By the same reasoning, the reverse transformations are given by

$$\left.\begin{array}{rcl} r &=& \sqrt{s^2+z^2} = \sqrt{x^2+y^2+z^2} \\ \theta &=& \arctan\left(\frac{s}{z}\right) = \arctan\left(\frac{\sqrt{x^2+y^2}}{z}\right) \\ \phi &=& \arctan\left(\frac{y}{x}\right) \end{array}\right\}. \tag{1.3.3}$$

We can also determine the unit vectors in spherical coordinates just as we did with cylindrical coordinates in Section 1.2. However, the order in which we list our coordinates is important. As a standard, the scientific community has decided all coordinate systems are to be *right-handed* meaning they obey the right-hand rule in the context of a cross product (e.g. $\hat{x} \times \hat{y} = \hat{z}$). Based on the direction in which we measure θ, it must be listed prior to ϕ because $\hat{r} \times \hat{\theta} = \hat{\phi}$. Therefore, a point in spherical coordinates would be given by the coordinate triplet (r, θ, ϕ).

Now using Eq. 1.1.1 along with $\hat{r} = \vec{r}/r$ and the fact that $\hat{\theta}$ must be perpendicular to $\hat{r}$ (initially with a positive s-component), the spherical unit vectors can be written as

$$\left.\begin{array}{rcl} \hat{r} &=& \sin\theta\,\hat{s} + \cos\theta\,\hat{z} \\ \hat{\theta} &=& \cos\theta\,\hat{s} - \sin\theta\,\hat{z} \\ \hat{\phi} &=& \hat{\phi} \end{array}\right\} \tag{1.3.4}$$

and

$$\left.\begin{array}{rcl} \hat{r} &=& \sin\theta\cos\phi\,\hat{x} + \sin\theta\sin\phi\,\hat{y} + \cos\theta\,\hat{z} \\ \hat{\theta} &=& \cos\theta\cos\phi\,\hat{x} + \cos\theta\sin\phi\,\hat{y} - \sin\theta\,\hat{z} \\ \hat{\phi} &=& -\sin\phi\,\hat{x} + \cos\phi\,\hat{y} \end{array}\right\}. \tag{1.3.5}$$

By the matrix method shown in Section 1.2, we can also write the Cartesian and cylindrical coordinates in terms of the spherical ones. They will be

$$\left.\begin{array}{rcl} \hat{s} &=& \sin\theta\,\hat{r} + \cos\theta\,\hat{\theta} \\ \hat{\phi} &=& \hat{\phi} \\ \hat{z} &=& \cos\theta\,\hat{r} - \sin\theta\,\hat{\theta} \end{array}\right\} \tag{1.3.6}$$

and

$$\left\{\begin{array}{rcl}\hat{x} &=& \sin\theta\cos\phi\,\hat{r} + \cos\theta\cos\phi\,\hat{\theta} - \sin\phi\,\hat{\phi} \\ \hat{y} &=& \sin\theta\sin\phi\,\hat{r} + \cos\theta\sin\phi\,\hat{\theta} + \cos\phi\,\hat{\phi} \\ \hat{z} &=& \cos\theta\,\hat{r} - \sin\theta\,\hat{\theta}\end{array}\right\}. \tag{1.3.7}$$

If you go through the matrix algebra as I have for all of these, you'll notice a pattern. Say for the sake of discussion our coefficient matrix is given as A. The pattern we will see is that the inverse matrix is equal to the transpose of the matrix (i.e. $A^{-1} = A^T$), where a transpose is simply a flip over the diagonal. This is not true for all matrices by any stretch, but it is true of **orthonormal** matrices, which are matrices formed by an orthonormal basis. This is exactly what we have here because the set of unit vectors for a coordinate system (e.g. $\{\hat{x}, \hat{y}, \hat{z}\}$) is referred to as a **basis** and should always be orthonormal. This makes finding inverse coordinate transformations very straightforward.

1.4 Bipolar and Elliptic

There are many more exotic options available, many of which are highly specialized by application. Some very interesting examples are bipolar and elliptic coordinates. Both of their names accurately suggest their nature. They both have essentially two origins positioned at $-a$ and $+a$ along the Cartesian x-axis and they are both defined by just two angles. This means they're both curvilinear in the plane with both sets of grid lines curved (i.e. no straight lines).

Position in bipolar coordinates is given by (τ, σ) with transformations given by

$$\left\{\begin{array}{rcl}x &=& a\dfrac{\sinh\tau}{\cosh\tau - \cos\sigma} \\ y &=& a\dfrac{\sin\sigma}{\cosh\tau - \cos\sigma}\end{array}\right\}. \tag{1.4.1}$$

If we define $\vec{r}_1$ and $\vec{r}_2$ as the position vectors relative to the origins at $x = -a$

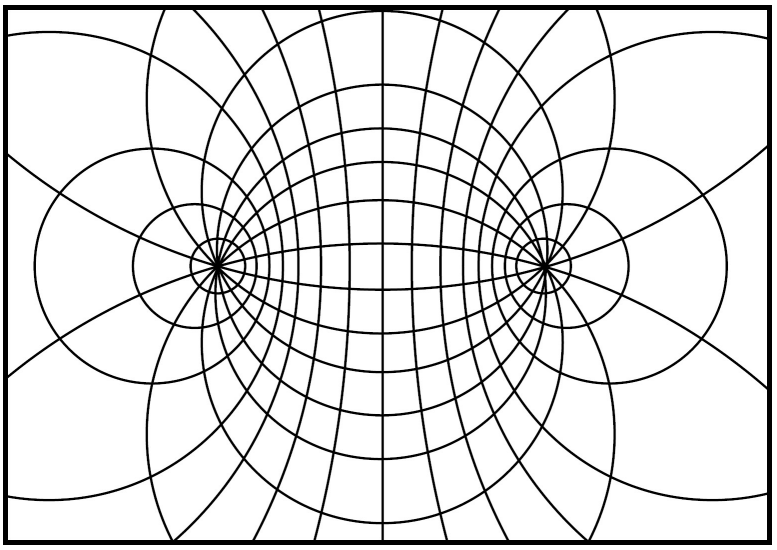


Figure 1.6: This is bipolar coordinates. The circles that intersect the origins at $\pm a$ along the horizontal axis are of constant σ and the circles that do not intersect at all are of constant τ.

and $x = +a$, respectively, then we can say

$$\left\{ \begin{array}{rcl} \tau & = & \ln\left(\dfrac{r_1}{r_2}\right) \\ \sigma & = & \arccos\left(\dfrac{\vec{r}_1 \bullet \vec{r}_2}{r_1 r_2}\right) \end{array} \right\}. \tag{1.4.2}$$

Position in elliptic coordinates is given by (μ, ν) with transformations given by

$$\left\{ \begin{array}{rcl} x & = & a\cosh\mu\cos\nu \\ y & = & a\sinh\mu\sin\nu \end{array} \right\}. \tag{1.4.3}$$

We can see from Eq. 1.4.3 that

$$\left(\frac{x}{a\cosh\mu}\right)^2 + \left(\frac{y}{a\sinh\mu}\right)^2 = \cos^2\nu + \sin^2\nu = 1 \tag{1.4.4}$$

matches the equation for an ellipse for constant μ. Also,

$$\left(\frac{x}{a\cos\nu}\right)^2 - \left(\frac{y}{a\sin\nu}\right)^2 = \cosh^2\mu - \sinh^2\mu = 1 \tag{1.4.5}$$

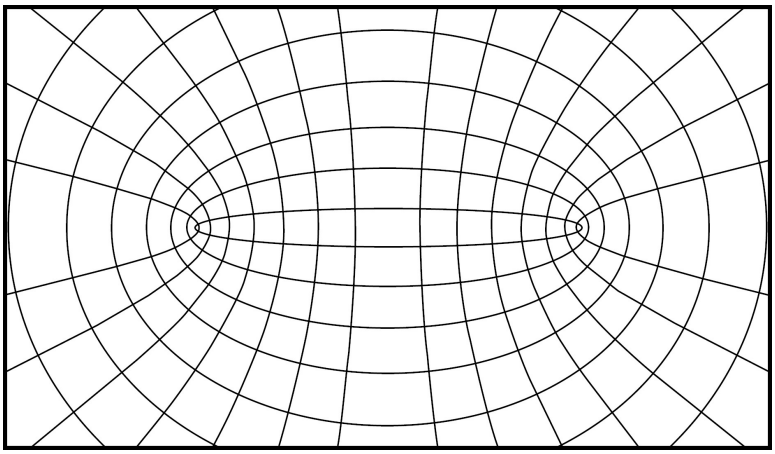


Figure 1.7: This is elliptic coordinates. The ellipses are of constant μ and the hyperbolas are of constant ν. The points $\pm a$ along the horizontal axis represent the foci of both the ellipses and hyperbolas.

matches the equation for a hyperbola for constant ν. A similar process can be used to show the grid lines in bipolar are all circles, but that derivation is much more algebraically and trigonometrically involved.

These two planar coordinate systems can be expanded into a wide variety of three-dimensional systems. We can project the grids along the z-axis to form bipolar cylindrical and elliptic cylindrical coordinates. They can be rotated about various axes to form toroidal, bispherical, oblate spheroidal, and prolate spheroidal coordinates. We can even take elliptic coordinates to the next dimension with its own angle definition resulting in ellipsoidal coordinates.

Chapter 2

Vector Algebra

2.1 Operators

The concept of operators is first introduced to students as children with basic arithmetic. We learn to add, subtract, multiply, and divide numbers. As mathematics progresses, we learn about exponents, parentheses, and the **order of operations** (PEMDAS) where we must use operators in a particular order. When we start learning algebra, we see that for every operator there is another operator that undoes it (i.e. an inverse operator like subtraction is for addition). It is at this point that the nature of operators sometimes becomes understated. Teachers will introduce the idea of functions on a basic level which can tend to sweep operators under the rug so to speak.

It isn't until some of us take classes in college like Abstract Algebra (or something similar) where we're reintroduced to operators. The arithmetic we've been doing all our lives is summed up in an algebraic structure known as a **field** (not to be confused with quantities we see in physics like the electric field). A mathematical field is a set of numbers closed under two operators. In the case of basic arithmetic, these two operations are addition and multiplication and the set of number is the real numbers. We'd write this as $(\mathbb{R}, +, *)$.

The other operations (e.g. exponents, subtraction, and division) are included through properties of fields such as inverses to maintain generality. For example, rather than subtract, we add an **additive inverse.** (e.g. $2 - 3 = 2 + (-3)$ where -3 is also in the set $\mathbb{R}$). Under higher levels of algebra we have multiplication and division, exponents and logs, sine and

arcsine, etc. In basic algebra they usually refer to sine or log as functions, but in reality they operate on one function (or number) to make another. All these operators obey certain properties. For example, operators in $(\mathbb{R}, +, *)$ obey the following properties:

- *Additive Identity*:
  For $a \in \mathbb{R}$, $a + 0 = a$.
- *Additive Inverse*:
  For $a \in \mathbb{R}$, $a + (-a) = 0$.
- *Multiplicative Identity*:
  For $a \in \mathbb{R}$, $a * 1 = a$.
- *Multiplicative Inverse*:
  For $a \in \mathbb{R}$, $a * a^{-1} = 1$.
- *Associative Property*:
  For $a, b, c \in \mathbb{R}$, $a + (b + c) = (a + b) + c$ and $a * (b * c) = (a * b) * c$.
- *Commutative Property*:
  For $a, b \in \mathbb{R}$, $a + b = b + a$ and $a * b = b * a$.
- *Distributive Property*:
  For $a, b, c \in \mathbb{R}$, $a * (b + c) = a * b + a * c$.

The properties listed above don't apply to all algebras. Matrices, for example, are not commutative under multiplication.

2.2 Vector Operators

In Section 2.1, we saw some of the properties associated with operating on scalar quantities and functions. Things happen a little differently when dealing with vector quantities and functions. The difference arises due to the directional nature of vectors, but we can still do our best to stick to the same terminology we used for scalars.

We can still write an additive statement for vectors like $\vec{A} + \vec{B} = \vec{C}$ so long as we break up the vectors into their components and add them separately. Adding the components separately retains the directional information through the operation. Subtraction of vectors is done in a similar fashion, by

taking advantage of the additive inverse. We can say $\vec{A}-\vec{B}=\vec{A}+(-\vec{B})=\vec{C}$. The only condition is that $-\vec{B}$ be in the vector space, which is very common in physical systems.

The division operator doesn't exist just as with matrices (in fact, we can write vectors as matrices because the operations are very similar), so to perform algebra usually suited for a division we need to be a little creative (e.g. with matrices, we would instead multiply by the multiplicative inverse). Multiplication does exist for vectors, but there are actually two types of multiplication: the dot product and the cross product. The necessity of this becomes clear when we consider the directional nature of vectors.

Through multiplication, the vectors will operate on each other. As this happens, it will be important how one vector is oriented relative to the other. If parallel components operate, then we have the dot product in which we lose directional information. This makes sense because if the components are operating parallel, then it's not really important in what direction this occurs. On the other hand, if orthogonal (perpendicular) components operate, then we have the cross product in which directional information is retained. This also makes sense, because information about the plane in which the vectors are operating will be important. Every plane has an orientation represented by a vector orthogonal to that plane, hence the cross product returns a vector orthogonal to both of the operating vectors. For vectors $\vec{A}$ and $\vec{B}$, we have the following definitions.

- *Dot Product* (by geometry):

$$\vec{A}\bullet\vec{B}=AB\cos\theta \tag{2.2.1}$$

  where θ is the angle between $\vec{A}$ and $\vec{B}$. Since $\cos(90^\circ)=0$, we see the dot product of orthogonal vectors gives a zero result. Also, since $\cos(0)=1$, we see the dot product of parallel vectors gives the maximum result.

- *Dot Product* (by components):

$$\begin{aligned}\vec{A}\bullet\vec{B} &= (A_x\hat{x}+A_y\hat{y}+A_z\hat{z})\bullet(B_x\hat{x}+B_y\hat{y}+B_z\hat{z})\\ &= A_xB_x\hat{x}\bullet\hat{x}+A_xB_y\hat{x}\bullet\hat{y}+A_xB_z\hat{x}\bullet\hat{z}\\ &\quad +A_yB_x\hat{y}\bullet\hat{x}+A_yB_y\hat{y}\bullet\hat{y}+A_yB_z\hat{y}\bullet\hat{z}\\ &\quad +A_zB_x\hat{z}\bullet\hat{x}+A_zB_y\hat{z}\bullet\hat{y}+A_zB_z\hat{z}\bullet\hat{z}\\ &= A_xB_x+A_yB_y+A_zB_z\end{aligned}$$

where we have taken advantage of Eq. 2.2.1 on the unit vectors (having a magnitude of one, by definition). We can write this more generally as

$$\vec{A} \bullet \vec{B} = \sum_{i=1}^{n} A_i B_i \tag{2.2.2}$$

where n represents the number of orthonormal components (usually 3 because it represents the number of dimensions).

- *Cross Product* (by geometry):

$$\vec{A} \times \vec{B} = AB \sin\theta \, \hat{n} \tag{2.2.3}$$

where θ is the angle between $\vec{A}$ and $\vec{B}$ and $\hat{n}$ is the unit vector orthogonal to both $\vec{A}$ and $\vec{B}$. Since $\sin(0) = 0$, we see the cross product of parallel vectors gives a zero result. Also, since $\sin(90°) = 1$, we see the cross product of orthogonal vectors gives the maximum result.

- *Cross Product* (by components):

$$\begin{aligned}
\vec{A} \times \vec{B} &= (A_x\hat{x} + A_y\hat{y} + A_z\hat{z}) \times (B_x\hat{x} + B_y\hat{y} + B_z\hat{z}) \\
&= A_xB_x\hat{x} \times \hat{x} + A_xB_y\hat{x} \times \hat{y} + A_xB_z\hat{x} \times \hat{z} \\
&\quad + A_yB_x\hat{y} \times \hat{x} + A_yB_y\hat{y} \times \hat{y} + A_yB_z\hat{y} \times \hat{z} \\
&\quad + A_zB_x\hat{z} \times \hat{x} + A_zB_y\hat{z} \times \hat{y} + A_zB_z\hat{z} \times \hat{z} \\
&= A_xB_y\hat{z} + A_xB_z\,(-\hat{y}) + A_yB_x\,(-\hat{z}) \\
&\quad + A_yB_z\hat{x} + A_zB_x\hat{y} + A_zB_y\,(-\hat{x})
\end{aligned}$$

$$\vec{A} \times \vec{B} = (A_yB_z - A_zB_y)\,\hat{x} + (A_zB_x - A_xB_z)\,\hat{y} + (A_xB_y - A_yB_x)\,\hat{z}$$

where we have taken advantage of Eq. 2.2.3 on the unit vectors (having a magnitude of one, by definition) noting that all our coordinate systems are right-handed (i.e. $\hat{x} \times \hat{y} = \hat{k}$ but $\hat{y} \times \hat{x} = -\hat{k}$). We can write this more simply as

$$\vec{A} \times \vec{B} = \det \begin{bmatrix} \hat{x} & \hat{y} & \hat{z} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{bmatrix} \tag{2.2.4}$$

which can be easily generalized for more dimensions if necessary.

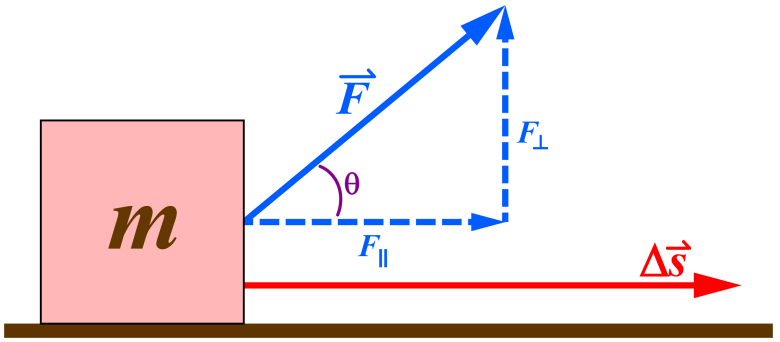


Figure 2.1: This diagram shows a constant force, $\vec{F}$, (with components labeled) acting on a mass, m, affecting a displacement, $\Delta\vec{s}$.

Example 2.2.1

Let's consider our definition of work: "Work is done on a body by a force over some displacement if that force directly affects the displacement of the body." For simplicity, we'll consider a force which is constant over the displacement. We can see clearly in Figure 2.1 that only the component of the force parallel to the displacement will affect the displacement. Taking advantage of Eq. 2.2.1, we have

$$\begin{aligned} W &= F_{||}\Delta s = (F\cos\theta)\,\Delta s = F\Delta s\cos\theta \\ W &= \vec{F}\bullet\Delta\vec{s}. \end{aligned}$$

This is the mathematical definition of the work done on a body by a constant force. Therefore, it makes perfect sense the dot product would be the operation to use in such a scenario. It may be confusing as to why we're multiplying these vector quantities in the first place. Well, we know these vectors must operate on each other if we're going to consider what they physically do together. Furthermore, we cannot add or subtract them because they're not like quantities (i.e. they don't have the same units) and there isn't a division operator for vectors. By process of elimination, this leaves only multiplication.

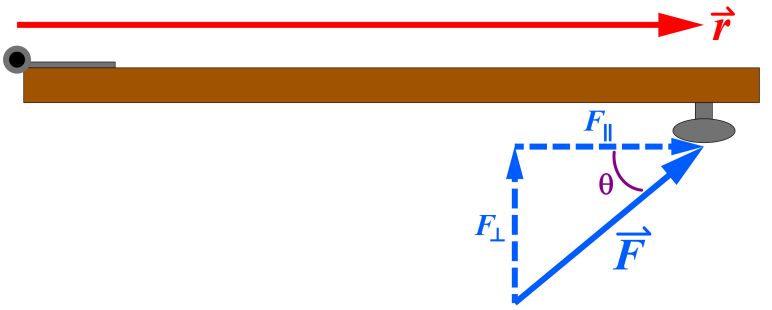


Figure 2.2: This diagram shows a constant force, $\vec{F}$, (with components labeled) acting on a door knob with a lever arm, $\vec{r}$, labeled.

Example 2.2.2

Let's consider a basic scenario: A constant force acting on a door knob. We can see clearly in Figure 2.2 that only the component of the force perpendicular to the door will generate a torque because it's the only one that can generate rotation. Taking advantage of Eq. 2.2.3, we have

$$\begin{aligned} \tau &= rF_{\perp} = r\left(F\sin\theta\right) = rF\sin\theta \\ \vec{\tau} &= \vec{r}\times\vec{F}. \end{aligned}$$

This is the mathematical definition of the torque on a body by a constant force at position $\vec{r}$. Therefore, it makes perfect sense the cross product would be the operation to use in such a scenario. Furthermore, just as with work done in Example 2.2.1, the only operation available to us is multiplication.

The dot product and cross product also obey properties similar to those found for real numbers and derivatives. For vectors $\vec{A}$, $\vec{B}$, and $\vec{C}$; and constant c;

- *Constant Multiple Properties*:

$$c\left(\vec{A}\bullet\vec{B}\right) = \left(c\vec{A}\right)\bullet\vec{B} = \vec{A}\bullet\left(c\vec{B}\right) \tag{2.2.5}$$

$$c\left(\vec{A}\times\vec{B}\right) = \left(c\vec{A}\right)\times\vec{B} = \vec{A}\times\left(c\vec{B}\right). \tag{2.2.6}$$

- *Distributive Properties*:

$$\vec{A} \bullet \left(\vec{B} + \vec{C}\right) = \vec{A} \bullet \vec{B} + \vec{A} \bullet \vec{C} \tag{2.2.7}$$

$$\vec{A} \times \left(\vec{B} + \vec{C}\right) = \vec{A} \times \vec{B} + \vec{A} \times \vec{C}. \tag{2.2.8}$$

- *Commutative Properties*:

$$\vec{A} \bullet \vec{B} = \vec{B} \bullet \vec{A} \tag{2.2.9}$$

$$\vec{A} \times \vec{B} = -\vec{B} \times \vec{A}. \tag{2.2.10}$$

  Note: The cross product changes sign.

- *Triple Product Rules*:

$$\vec{A} \bullet \left(\vec{B} \times \vec{C}\right) = \vec{C} \bullet \left(\vec{A} \times \vec{B}\right) = \vec{B} \bullet \left(\vec{C} \times \vec{A}\right) \tag{2.2.11}$$

$$\vec{A} \times \left(\vec{B} \times \vec{C}\right) = \vec{B}\left(\vec{A} \bullet \vec{C}\right) - \vec{C}\left(\vec{A} \bullet \vec{B}\right) \tag{2.2.12}$$

It should be stated explicitly here that neither the dot product nor the cross product is associative. That means, when writing triple products, parentheses must always be present.

Chapter 3

Vector Calculus

3.1 Calculus

Early on in calculus, we're shown how to take a derivative of a function. Then later, we see the integral (also called an anti-derivative). These are both operators and they obey certain properties just like the algebraic operators from Section 2.1. For real-valued functions $f(x)$ and $g(x)$ and real-valued constant c,

- *Fundamental Theorem of Calculus (or Inverse Property)*:

$$\int_a^b \frac{d}{dx}(f)\,dx = \int_a^b df = f|_{x=b} - f|_{x=a}\,. \tag{3.1.1}$$

- *Chain Rule*:

$$\frac{d}{dx}(f) = \frac{d}{du}(f)\frac{du}{dx}. \tag{3.1.2}$$

- *Constant Multiple Property*:

$$c\frac{d}{dx}(f) = \frac{d}{dx}(cf)\,. \tag{3.1.3}$$

- *Distributive Property*:

$$\frac{d}{dx}(f+g) = \frac{d}{dx}(f) + \frac{d}{dx}(g)\,. \tag{3.1.4}$$

- *Product Rule*:

$$\frac{d}{dx}(f * g) = \frac{d}{dx}(f) * g + f * \frac{d}{dx}(g). \tag{3.1.5}$$

- *Quotient Rule*:

$$\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{\frac{d}{dx}(f) * g - f * \frac{d}{dx}(g)}{g^2}. \tag{3.1.6}$$

However, I find listing the derivative quotient rule to be redundant because we can simply apply the product rule to $f * g^{-1}$ where the negative one exponent represents the multiplicative inverse, not the inverse function. The derivative product rule could also be referred to as a distributive property over multiplication, but I think the name would tempt those new to the idea to distribute the derivative operator just like we do for addition, so we'll just call it the product rule to retain clarity.

3.2 Del Operator

In Section 2.1, we expanded our knowledge of operators in general as well as emphasized the importance of operators in mathematics. In Section 2.2, we were exposed to how our algebraic operators behaved with vectors ...so what about the calculus operators from Section 3.1? Can they apply to vectors?

Vectors (and scalars for that matter) can be functions of both space and time. That's four variables! This means we'll be dealing with partial derivatives rather than total derivatives. In Cartesian coordinates (see Section 1.1), we have the following options:

$$\left\{\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right\}$$

For the time being, we'll keep time separate and only consider space. In space, vector functions involve direction, so a derivative operator for vectors should incorporate that as well. How about a vector with derivative components? We call it the **del operator** and we use the nabla symbol to represent it. In Cartesian coordinates, it takes the form

$$\vec{\nabla} \equiv \hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z} \tag{3.2.1}$$

where the unit vectors have been placed in front to avoid confusion. This seems simple enough, but how does it operate? The del operator can operate on both scalar and vector functions. When it operates on a scalar function, $f(x, y, z)$, we have

$$\vec{\nabla} f = \frac{\partial f}{\partial x}\hat{x} + \frac{\partial f}{\partial y}\hat{y} + \frac{\partial f}{\partial z}\hat{z}\,. \tag{3.2.2}$$

This is called the **gradient** and it measures how the scalar function, f, changes in space. This means the change of a scalar function is a vector, which makes sense. We would want to know in what direction it's changing the most.

However, if del operates on another vector, we have two options: dot product and cross product. Using Eqs. 2.2.2 and 2.2.4 with a vector field, $\vec{A}(x, y, z)$, results in

$$\vec{\nabla} \bullet \vec{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}\,. \tag{3.2.3}$$

$$\vec{\nabla} \times \vec{A} = \det \begin{bmatrix} \hat{x} & \hat{y} & \hat{z} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ A_x & A_y & A_z \end{bmatrix} \tag{3.2.4}$$

where, again, the dot product results in a scalar and the cross product a vector. The next question on everyone's minds: "Sure, but what do they mean?!"

Eq. 3.2.3 is called the **divergence** and it measures how a vector field, $\vec{A}$, diverges or spreads at a single position. In other words, if the divergence (at a point) is positive, then there are more vectors directed outward surrounding that point than there are directed inward. The opposite is true for a negative divergence. By the same reasoning, a divergence of zero implies there is the same amount outward as inward. This is sounding very abstract, I know, but we're keeping definitions as general as possible. This concept could apply to a multitude of situations (e.g. velocity or flow of fluids, electrodynamics, etc.).

Eq. 3.2.4 is called the **curl** and it measures how a vector field, $\vec{A}$, curls or circulates at a single position. This makes perfect sense if applied to fluid

flow, but what about something like electromagnetic fields where nothing is actually circulating? We can see some of these fields curl, but they certainly don't circulate, right? True, and the field itself bending doesn't always indicate a non-zero curl (a straight field doesn't indicate zero curl either). It's best to think of the curl in terms of how something would respond to the field when placed inside. In a circulating fluid, a small object might rotate or revolve. In an electric field, one charge would move toward or away from another. In a magnetic field, a moving charge will travel in circle-like paths. We can attain a visual based on how these foreign bodies move as a result of the field's influence. Furthermore, the direction of the angular velocity of the body will be in the same direction as the curl.

It should be noted that we're not really dotting or crossing two vectors together. Yes, the del operator has a vector form, but it's more an operator than a vector. We lose the commutative properties of the two products because del has to operate something (i.e. $\vec{\nabla} \bullet \vec{A} \neq \vec{A} \bullet \vec{\nabla}$). Because it doesn't obey all properties of vectors, the rigorous among us refuse to call del a vector.

We can also use the del operator to take a second derivative. However, since this operator is changing the nature of our function between scalar and vector and vice versa, we have several options mathematically: divergence of a gradient, gradient of a divergence, curl of a gradient, divergence of a curl, and curl of a curl. This might seem like a lot, but we can eliminate several of them. First, the divergence of the gradient, $\vec{\nabla} \bullet \left(\vec{\nabla} f\right)$, has a special name: the **Laplacian**. It is short-handed with the symbol $\vec{\nabla}^2$ and is represented in Cartesian coordinates by

$$\vec{\nabla}^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}. \tag{3.2.5}$$

The curl of a gradient and the divergence of a curl are both zero, which we can show mathematically as

$$\vec{\nabla} \times \left(\vec{\nabla} f\right) = 0 \tag{3.2.6}$$

and

$$\vec{\nabla} \bullet \left(\vec{\nabla} \times \vec{A}\right) = 0. \tag{3.2.7}$$

Both of these can be mathematically proven using Eqs. 3.2.2, 3.2.3, and 3.2.4 by realizing the partial derivatives are commutative:

$$\frac{\partial}{\partial x}\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}\frac{\partial f}{\partial x} \quad \Rightarrow \left(\frac{\partial}{\partial x}\frac{\partial f}{\partial y} - \frac{\partial}{\partial y}\frac{\partial f}{\partial x}\right) = 0.$$

The gradient of the divergence, $\vec{\nabla}\left(\vec{\nabla} \bullet \vec{A}\right)$, is not zero, but is extremely rare in physical systems. The curl of the curl obeys the identity

$$\vec{\nabla} \times \left(\vec{\nabla} \times \vec{A}\right) = \vec{\nabla}\left(\vec{\nabla} \bullet \vec{A}\right) - \vec{\nabla}^2\vec{A}, \tag{3.2.8}$$

which contains the gradient of the divergence and the Laplacian, second derivatives already seen. The Laplacian of a vector is defined in Cartesian coordinates as

$$\vec{\nabla}^2\vec{A} \equiv \left(\vec{\nabla}^2 A_x\right)\hat{x} + \left(\vec{\nabla}^2 A_y\right)\hat{y} + \left(\vec{\nabla}^2 A_z\right)\hat{z},$$

which is a vector with Laplacian components. As simple an extension as this might be for the Laplacian, you'll probably never need to write this out in a particular coordinate system anyway.

Just as like Eq. 3.1.5, there are similar product rules for the del operator. However, since del operates in three different ways on two different types of quantities, there are six product rules. For vector fields $\vec{A}(x,y,z)$ and $\vec{B}(x,y,z)$, and scalar functions $f(x,y,z)$ and $g(x,y,z)$, they are:

$$\vec{\nabla}(fg) = \left(\vec{\nabla}f\right)g + f\left(\vec{\nabla}g\right) \tag{3.2.9}$$

$$\vec{\nabla} \bullet \left(f\vec{A}\right) = \vec{A} \bullet \left(\vec{\nabla}f\right) + f\left(\vec{\nabla} \bullet \vec{A}\right) \tag{3.2.10}$$

$$\vec{\nabla} \times \left(f\vec{A}\right) = -\vec{A} \times \left(\vec{\nabla}f\right) + f\left(\vec{\nabla} \times \vec{A}\right) \tag{3.2.11}$$

$$\vec{\nabla} \bullet \left(\vec{A} \times \vec{B}\right) = \vec{B} \bullet \left(\vec{\nabla} \times \vec{A}\right) - \vec{A} \bullet \left(\vec{\nabla} \times \vec{B}\right) \tag{3.2.12}$$

$$\begin{aligned}\vec{\nabla}\left(\vec{A} \bullet \vec{B}\right) &= \vec{A} \times \left(\vec{\nabla} \times \vec{B}\right) + \vec{B} \times \left(\vec{\nabla} \times \vec{A}\right) \\ &\quad + \left(\vec{A} \bullet \vec{\nabla}\right)\vec{B} + \left(\vec{B} \bullet \vec{\nabla}\right)\vec{A}\end{aligned} \tag{3.2.13}$$

$$\begin{aligned}\vec{\nabla} \times \left(\vec{A} \times \vec{B}\right) &= \left(\vec{B} \bullet \vec{\nabla}\right)\vec{A} - \vec{B}\left(\vec{\nabla} \bullet \vec{A}\right) \\ &\quad - \left(\vec{A} \bullet \vec{\nabla}\right)\vec{B} + \vec{A}\left(\vec{\nabla} \bullet \vec{B}\right).\end{aligned} \tag{3.2.14}$$

3.3 Non-Cartesian Del Operators

Eqs. 3.2.6 through 3.2.14 made no reference to any coordinate system. These equations are true in all coordinate systems and so we call them del operator identities. However, we did quite a bit of work in Section 3.2 in Cartesian coordinates. If we want to write out the gradient, divergence, or curl in another coordinate system, then we'll need to transform the operators and the vector they're operating on. This will take a bit of finesse and the result won't always look so simple.

Example 3.3.1

The del operator, gradient, divergence, and curl can be found for *any* coordinate system by performing the following steps. For context, we'll find them for cylindrical coordinates (Section 1.2).

1. *Find the Cartesian variables in terms of the variables of the new coordinate system.*
   Eq. 1.2.1 ...check!

2. *Find the variables of the new coordinate system in terms of the Cartesian variables.*
   Eq. 1.2.2 ...check!

3. *Find the unit vectors in the new coordinate system in terms of the Cartesian unit vectors.*
   Eq. 1.2.3 ...check!

4. *Find the Cartesian unit vectors in terms of the unit vectors in the new coordinate system.*
   Eq. 1.2.7 ...check!

5. *Determine the cross product combinations of the new unit vectors using the right-hand rule.*
   Based on the order in which we've listed the variables, (s, ϕ, z), and Eq. 2.2.10, we can conclude

$$\left.\begin{array}{lll} \hat{s} \times \hat{\phi} = \hat{z} & \text{and} & \hat{\phi} \times \hat{s} = -\hat{z} \\ \hat{z} \times \hat{s} = \hat{\phi} & \text{and} & \hat{s} \times \hat{z} = -\hat{\phi} \\ \hat{\phi} \times \hat{z} = \hat{s} & \text{and} & \hat{z} \times \hat{\phi} = -\hat{s} \end{array}\right\}.$$

6. *Evaluate all the possible first derivatives of the new variables with respect to the Cartesian variables (there are 9 derivatives total) and then transform back to the new variables.*
Using Eqs. 1.2.2 and then 1.2.1, we see that

$$\frac{\partial s}{\partial z} = \frac{\partial \phi}{\partial z} = \frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} = \frac{\partial z}{\partial z} = 0,$$

$$\left\{ \begin{array}{l} \dfrac{\partial s}{\partial x} = \dfrac{x}{\sqrt{x^2+y^2}} = \dfrac{s\cos\phi}{s} = \cos\phi \\ \dfrac{\partial s}{\partial y} = \dfrac{y}{\sqrt{x^2+y^2}} = \dfrac{s\sin\phi}{s} = \sin\phi \end{array} \right\},$$

and

$$\left\{ \begin{array}{l} \dfrac{\partial \phi}{\partial x} = \dfrac{-y}{x^2+y^2} = \dfrac{-s\sin\phi}{s^2} = \dfrac{-\sin\phi}{s} \\ \dfrac{\partial \phi}{\partial y} = \dfrac{x}{x^2+y^2} = \dfrac{s\cos\phi}{s^2} = \dfrac{\cos\phi}{s} \end{array} \right\}.$$

7. *Evaluate all the possible first derivatives of the new unit vectors with respect to the new variables (there are 9 derivatives total).*
Unlike in Cartesian, the direction of cylindrical unit vectors depends on position in space, so this is necessary (and happens to be the source of most of our trouble). Using Eq. 1.2.3, we see that

$$\frac{\partial \hat{s}}{\partial s} = \frac{\partial \hat{s}}{\partial z} = \frac{\partial \hat{\phi}}{\partial s} = \frac{\partial \hat{\phi}}{\partial z} = \frac{\partial \hat{z}}{\partial s} = \frac{\partial \hat{z}}{\partial \phi} = \frac{\partial \hat{z}}{\partial z} = 0,$$

$$\frac{\partial \hat{s}}{\partial \phi} = -\sin\phi\,\hat{x} + \cos\phi\,\hat{y} = \hat{\phi},$$

and

$$\frac{\partial \hat{\phi}}{\partial \phi} = -\cos\phi\,\hat{x} - \sin\phi\,\hat{y} = -\hat{s}.$$

8. *Use the chain rule to expand each Cartesian derivative operator into the new coordinate operators.*
   By the chain rule (Eq. 3.1.2) generalized to multi-variable partial derivatives, we have

$$\left\{\begin{aligned}\frac{\partial}{\partial x} &= \frac{\partial s}{\partial x}\frac{\partial}{\partial s} + \frac{\partial \phi}{\partial x}\frac{\partial}{\partial \phi} + \frac{\partial z}{\partial x}\frac{\partial}{\partial z}\\ \frac{\partial}{\partial y} &= \frac{\partial s}{\partial y}\frac{\partial}{\partial s} + \frac{\partial \phi}{\partial y}\frac{\partial}{\partial \phi} + \frac{\partial z}{\partial y}\frac{\partial}{\partial z}\\ \frac{\partial}{\partial z} &= \frac{\partial s}{\partial z}\frac{\partial}{\partial s} + \frac{\partial \phi}{\partial z}\frac{\partial}{\partial \phi} + \frac{\partial z}{\partial z}\frac{\partial}{\partial z}\end{aligned}\right\}.$$

   Making substitutions from step 6, we get

$$\left\{\begin{aligned}\frac{\partial}{\partial x} &= \cos\phi\frac{\partial}{\partial s} - \frac{\sin\phi}{s}\frac{\partial}{\partial \phi}\\ \frac{\partial}{\partial y} &= \sin\phi\frac{\partial}{\partial s} + \frac{\cos\phi}{s}\frac{\partial}{\partial \phi}\\ \frac{\partial}{\partial z} &= \frac{\partial}{\partial z}\end{aligned}\right\}.$$

   It is now clear that the operator with respect to z remains unaffected, which makes sense given Eq. 1.2.1.

9. *Make substitutions from steps 4 and 8 into Eq. 3.2.1.*

$$\vec{\nabla} \equiv \hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z}$$

   Making substitutions from Eq. 1.2.7 and step 8, we get

$$\begin{aligned}\vec{\nabla} \;=\; &\left(\cos\phi\,\hat{s} - \sin\phi\,\hat{\phi}\right)\left(\cos\phi\frac{\partial}{\partial s} - \frac{\sin\phi}{s}\frac{\partial}{\partial \phi}\right)\\ &+ \left(\sin\phi\,\hat{s} + \cos\phi\,\hat{\phi}\right)\left(\sin\phi\frac{\partial}{\partial s} + \frac{\cos\phi}{s}\frac{\partial}{\partial \phi}\right)\\ &+ \hat{z}\frac{\partial}{\partial z}.\end{aligned}$$

We can now expand the two binomial products (i.e. using the distributive property of multiplication) arriving at

$$\begin{aligned}\vec{\nabla} \;=\; & \hat{s}\cos^2\phi\frac{\partial}{\partial s} - \hat{\phi}\sin\phi\cos\phi\frac{\partial}{\partial s} - \hat{s}\frac{\sin\phi\cos\phi}{s}\frac{\partial}{\partial\phi} + \hat{\phi}\frac{\sin^2\phi}{s}\frac{\partial}{\partial\phi} \\ & + \hat{s}\sin^2\phi\frac{\partial}{\partial s} + \hat{\phi}\sin\phi\cos\phi\frac{\partial}{\partial s} + \hat{s}\frac{\sin\phi\cos\phi}{s}\frac{\partial}{\partial\phi} + \hat{\phi}\frac{\cos^2\phi}{s}\frac{\partial}{\partial\phi} \\ & + \hat{z}\frac{\partial}{\partial z}.\end{aligned}$$

Several terms will cancel and the remaining terms can be simplified by $\sin^2\phi + \cos^2\phi = 1$ resulting in a del operator of

$$\boxed{\vec{\nabla} = \hat{s}\frac{\partial}{\partial s} + \hat{\phi}\frac{1}{s}\frac{\partial}{\partial\phi} + \hat{z}\frac{\partial}{\partial z}} \tag{3.3.1}$$

for cylindrical coordinates. This is close to what we might expect with the exception of the factor of $1/s$ in the $\hat{\phi}$ term.

10. *Operate del on an arbitrary scalar function $f(s,\phi,z)$ to find the gradient.*

$$\boxed{\vec{\nabla} f = \frac{\partial f}{\partial s}\hat{s} + \frac{1}{s}\frac{\partial f}{\partial\phi}\hat{\phi} + \frac{\partial f}{\partial z}\hat{z}}$$

11. *Operate del on an arbitrary vector field $\vec{A}(s,\phi,z)$ using the dot product.* Using the dot product, we get

$$\vec{\nabla}\bullet\vec{A} = \left[\hat{s}\frac{\partial}{\partial s} + \hat{\phi}\frac{1}{s}\frac{\partial}{\partial\phi} + \hat{z}\frac{\partial}{\partial z}\right]\bullet\vec{A}.$$

However, this is where we have to be very careful about what we mean by the del operator. As stated in Section 3.2, del is an operator before it's a vector. We didn't have to worry about this in the Cartesian case because the unit vectors had constant direction. In cylindrical coordinates, this is no longer true, so we must make sure del operates ***before*** we perform the dot product. Taking great care to not accidentally commute any terms, we get

$$\vec{\nabla}\bullet\vec{A} = \hat{s}\bullet\frac{\partial\vec{A}}{\partial s} + \hat{\phi}\bullet\frac{1}{s}\frac{\partial\vec{A}}{\partial\phi} + \hat{z}\bullet\frac{\partial\vec{A}}{\partial z}.$$

Writing the vector field in terms of unit vectors as $\vec{A} = A_s\hat{s} + A_\phi\hat{\phi} + A_z\hat{z}$, we get

$$\begin{aligned}\vec{\nabla} \bullet \vec{A} &= \hat{s} \bullet \frac{\partial}{\partial s}\left(A_s\hat{s} + A_\phi\hat{\phi} + A_z\hat{z}\right) \\ &\quad + \hat{\phi} \bullet \frac{1}{s}\frac{\partial}{\partial \phi}\left(A_s\hat{s} + A_\phi\hat{\phi} + A_z\hat{z}\right) \\ &\quad + \hat{z} \bullet \frac{\partial}{\partial z}\left(A_s\hat{s} + A_\phi\hat{\phi} + A_z\hat{z}\right).\end{aligned}$$

We can now distribute the derivative operators and perform the necessary product rules (Eq. 3.1.5) resulting in

$$\begin{aligned}\vec{\nabla} \bullet \vec{A} &= \hat{s} \bullet \left[\frac{\partial}{\partial s}\left(A_s\hat{s}\right) + \frac{\partial}{\partial s}\left(A_\phi\hat{\phi}\right) + \frac{\partial}{\partial s}\left(A_z\hat{z}\right)\right] \\ &\quad + \hat{\phi} \bullet \frac{1}{s}\left[\frac{\partial}{\partial \phi}\left(A_s\hat{s}\right) + \frac{\partial}{\partial \phi}\left(A_\phi\hat{\phi}\right) + \frac{\partial}{\partial \phi}\left(A_z\hat{z}\right)\right] \\ &\quad + \hat{z} \bullet \left[\frac{\partial}{\partial z}\left(A_s\hat{s}\right) + \frac{\partial}{\partial z}\left(A_\phi\hat{\phi}\right) + \frac{\partial}{\partial z}\left(A_z\hat{z}\right)\right]\end{aligned}$$

$$\begin{aligned}\vec{\nabla} \bullet \vec{A} &= \hat{s} \bullet \left[\frac{\partial A_s}{\partial s}\hat{s} + A_s\frac{\partial \hat{s}}{\partial s} + \frac{\partial A_\phi}{\partial s}\hat{\phi} + A_\phi\frac{\partial \hat{\phi}}{\partial s} + \frac{\partial A_z}{\partial s}\hat{z} + A_z\frac{\partial \hat{z}}{\partial s}\right] \\ &\quad + \hat{\phi} \bullet \frac{1}{s}\left[\frac{\partial A_s}{\partial \phi}\hat{s} + A_s\frac{\partial \hat{s}}{\partial \phi} + \frac{\partial A_\phi}{\partial \phi}\hat{\phi} + A_\phi\frac{\partial \hat{\phi}}{\partial \phi} + \frac{\partial A_z}{\partial \phi}\hat{z} + A_z\frac{\partial \hat{z}}{\partial \phi}\right] \\ &\quad + \hat{z} \bullet \left[\frac{\partial A_s}{\partial z}\hat{s} + A_s\frac{\partial \hat{s}}{\partial z} + \frac{\partial A_\phi}{\partial z}\hat{\phi} + A_\phi\frac{\partial \hat{\phi}}{\partial z} + \frac{\partial A_z}{\partial z}\hat{z} + A_z\frac{\partial \hat{z}}{\partial z}\right].\end{aligned}$$

Making substitutions from step 7, we get

$$\begin{aligned}\vec{\nabla} \bullet \vec{A} &= \hat{s} \bullet \left[\frac{\partial A_s}{\partial s}\hat{s} + \frac{\partial A_\phi}{\partial s}\hat{\phi} + \frac{\partial A_z}{\partial s}\hat{z}\right] \\ &\quad + \hat{\phi} \bullet \frac{1}{s}\left[\frac{\partial A_s}{\partial \phi}\hat{s} + A_s\hat{\phi} + \frac{\partial A_\phi}{\partial \phi}\hat{\phi} + A_\phi\left(-\hat{s}\right) + \frac{\partial A_z}{\partial \phi}\hat{z}\right] \\ &\quad + \hat{z} \bullet \left[\frac{\partial A_s}{\partial z}\hat{s} + \frac{\partial A_\phi}{\partial z}\hat{\phi} + \frac{\partial A_z}{\partial z}\hat{z}\right].\end{aligned}$$

Finally, we can operate with the dot product, which results in

$$\begin{aligned}\vec{\nabla} \bullet \vec{A} &= \frac{\partial A_s}{\partial s} + \frac{1}{s}\left[A_s + \frac{\partial A_\phi}{\partial \phi}\right] + \frac{\partial A_z}{\partial z} \\ &= \frac{\partial A_s}{\partial s} + \frac{1}{s}A_s + \frac{1}{s}\frac{\partial A_\phi}{\partial \phi} + \frac{\partial A_z}{\partial z} \\ &= \frac{1}{s}\left[s\frac{\partial A_s}{\partial s} + (1)\, A_s\right] + \frac{1}{s}\frac{\partial A_\phi}{\partial \phi} + \frac{\partial A_z}{\partial z}.\end{aligned}$$

Since $\partial s/\partial s = 1$, we can perform something I like to call *voodoo math* (with a little foresight; we can add zeros, multiply by ones, add and subtract constants, etc. to simplify a mathematical expression) and we have

$$\vec{\nabla} \bullet \vec{A} = \frac{1}{s}\left[s\frac{\partial A_s}{\partial s} + \frac{\partial s}{\partial s}A_s\right] + \frac{1}{s}\frac{\partial A_\phi}{\partial \phi} + \frac{\partial A_z}{\partial z}$$

and we can see the quantity in brackets matches the form of Eq. 3.1.5. Rewriting, we arrive at our final answer of

$$\boxed{\vec{\nabla} \bullet \vec{A} = \frac{1}{s}\frac{\partial}{\partial s}\left(sA_s\right) + \frac{1}{s}\frac{\partial A_\phi}{\partial \phi} + \frac{\partial A_z}{\partial z}}$$

12. *Operate del on an arbitrary vector field $\vec{A}(s, \phi, z)$ using the cross product.* Using the cross product, we get

$$\vec{\nabla} \times \vec{A} = \left[\hat{s}\frac{\partial}{\partial s} + \hat{\phi}\frac{1}{s}\frac{\partial}{\partial \phi} + \hat{z}\frac{\partial}{\partial z}\right] \times \vec{A}.$$

However, this is where we have to be very careful about what we mean by the del operator. As stated in Section 3.2, del is an operator before it's a vector. We didn't have to worry about this in the Cartesian case because the unit vectors had constant direction. In cylindrical coordinates, this is no longer true, so we must make sure del operates ***before*** we perform the cross product. Taking great care to not accidentally commute any terms, we get

$$\vec{\nabla} \times \vec{A} = \hat{s} \times \frac{\partial \vec{A}}{\partial s} + \hat{\phi} \times \frac{1}{s}\frac{\partial \vec{A}}{\partial \phi} + \hat{z} \times \frac{\partial \vec{A}}{\partial z}.$$

Writing the vector field in terms of unit vectors as $\vec{A} = A_s\hat{s} + A_\phi\hat{\phi} + A_z\hat{z}$, we get

$$\begin{aligned}\vec{\nabla}\times\vec{A} &= \hat{s}\times\frac{\partial}{\partial s}\left(A_s\hat{s} + A_\phi\hat{\phi} + A_z\hat{z}\right)\\ &\quad +\hat{\phi}\times\frac{1}{s}\frac{\partial}{\partial\phi}\left(A_s\hat{s} + A_\phi\hat{\phi} + A_z\hat{z}\right)\\ &\quad +\hat{z}\times\frac{\partial}{\partial z}\left(A_s\hat{s} + A_\phi\hat{\phi} + A_z\hat{z}\right).\end{aligned}$$

We can now distribute the derivative operators and perform the necessary product rules (Eq. 3.1.5) resulting in

$$\begin{aligned}\vec{\nabla}\times\vec{A} &= \hat{s}\times\left[\frac{\partial}{\partial s}\left(A_s\hat{s}\right) + \frac{\partial}{\partial s}\left(A_\phi\hat{\phi}\right) + \frac{\partial}{\partial s}\left(A_z\hat{z}\right)\right]\\ &\quad +\hat{\phi}\times\frac{1}{s}\left[\frac{\partial}{\partial\phi}\left(A_s\hat{s}\right) + \frac{\partial}{\partial\phi}\left(A_\phi\hat{\phi}\right) + \frac{\partial}{\partial\phi}\left(A_z\hat{z}\right)\right]\\ &\quad +\hat{z}\times\left[\frac{\partial}{\partial z}\left(A_s\hat{s}\right) + \frac{\partial}{\partial z}\left(A_\phi\hat{\phi}\right) + \frac{\partial}{\partial z}\left(A_z\hat{z}\right)\right]\end{aligned}$$

$$\begin{aligned}\vec{\nabla}\times\vec{A} &= \hat{s}\times\left[\frac{\partial A_s}{\partial s}\hat{s} + A_s\frac{\partial\hat{s}}{\partial s} + \frac{\partial A_\phi}{\partial s}\hat{\phi} + A_\phi\frac{\partial\hat{\phi}}{\partial s} + \frac{\partial A_z}{\partial s}\hat{z} + A_z\frac{\partial\hat{z}}{\partial s}\right]\\ &\quad +\hat{\phi}\times\frac{1}{s}\left[\frac{\partial A_s}{\partial\phi}\hat{s} + A_s\frac{\partial\hat{s}}{\partial\phi} + \frac{\partial A_\phi}{\partial\phi}\hat{\phi} + A_\phi\frac{\partial\hat{\phi}}{\partial\phi} + \frac{\partial A_z}{\partial\phi}\hat{z} + A_z\frac{\partial\hat{z}}{\partial\phi}\right]\\ &\quad +\hat{z}\times\left[\frac{\partial A_s}{\partial z}\hat{s} + A_s\frac{\partial\hat{s}}{\partial z} + \frac{\partial A_\phi}{\partial z}\hat{\phi} + A_\phi\frac{\partial\hat{\phi}}{\partial z} + \frac{\partial A_z}{\partial z}\hat{z} + A_z\frac{\partial\hat{z}}{\partial z}\right].\end{aligned}$$

Making substitutions from step 7, we get

$$\begin{aligned}\vec{\nabla}\times\vec{A} &= \hat{s}\times\left[\frac{\partial A_s}{\partial s}\hat{s} + \frac{\partial A_\phi}{\partial s}\hat{\phi} + \frac{\partial A_z}{\partial s}\hat{z}\right]\\ &\quad +\hat{\phi}\times\frac{1}{s}\left[\frac{\partial A_s}{\partial\phi}\hat{s} + A_s\hat{\phi} + \frac{\partial A_\phi}{\partial\phi}\hat{\phi} + A_\phi\left(-\hat{s}\right) + \frac{\partial A_z}{\partial\phi}\hat{z}\right]\\ &\quad +\hat{z}\times\left[\frac{\partial A_s}{\partial z}\hat{s} + \frac{\partial A_\phi}{\partial z}\hat{\phi} + \frac{\partial A_z}{\partial z}\hat{z}\right].\end{aligned}$$

Finally, we can operate with the cross product taking advantage of the relationships we found in step 5, which results in

$$\begin{aligned}\vec{\nabla}\times\vec{A} &= \frac{\partial A_\phi}{\partial s}(+\hat{z})+\frac{\partial A_z}{\partial s}\left(-\hat{\phi}\right)\\ &\quad+\frac{1}{s}\left[\frac{\partial A_s}{\partial \phi}(-\hat{z})-A_\phi(-\hat{z})+\frac{\partial A_z}{\partial \phi}(+\hat{s})\right]\\ &\quad+\frac{\partial A_s}{\partial z}\left(+\hat{\phi}\right)+\frac{\partial A_\phi}{\partial z}(-\hat{s})\end{aligned}$$

$$\begin{aligned}\vec{\nabla}\times\vec{A} &= \frac{\partial A_\phi}{\partial s}\hat{z}-\frac{\partial A_z}{\partial s}\hat{\phi}-\frac{1}{s}\frac{\partial A_s}{\partial \phi}\hat{z}+\frac{1}{s}A_\phi\hat{z}\\ &\quad+\frac{1}{s}\frac{\partial A_z}{\partial \phi}\hat{s}+\frac{\partial A_s}{\partial z}\hat{\phi}-\frac{\partial A_\phi}{\partial z}\hat{s}.\end{aligned}$$

Now we can group terms of similar direction together arriving at

$$\begin{aligned}\vec{\nabla}\times\vec{A} &= \left[\frac{1}{s}\frac{\partial A_z}{\partial \phi}-\frac{\partial A_\phi}{\partial z}\right]\hat{s}+\left[\frac{\partial A_s}{\partial z}-\frac{\partial A_z}{\partial s}\right]\hat{\phi}\\ &\quad+\left[\frac{\partial A_\phi}{\partial s}+\frac{1}{s}A_\phi-\frac{1}{s}\frac{\partial A_s}{\partial \phi}\right]\hat{z}\end{aligned}$$

$$\begin{aligned}\vec{\nabla}\times\vec{A} &= \left[\frac{1}{s}\frac{\partial A_z}{\partial \phi}-\frac{\partial A_\phi}{\partial z}\right]\hat{s}+\left[\frac{\partial A_s}{\partial z}-\frac{\partial A_z}{\partial s}\right]\hat{\phi}\\ &\quad+\frac{1}{s}\left[s\frac{\partial A_\phi}{\partial s}+(1)\,A_\phi-\frac{\partial A_s}{\partial \phi}\right]\hat{z}.\end{aligned}$$

Since $\partial s/\partial s = 1$, we can perform something I like to call *voodoo math* (with a little foresight; we can add zeros, multiply by ones, add and subtract constants, etc. to simplify a mathematical expression) and we have

$$\begin{aligned}\vec{\nabla}\times\vec{A} &= \left[\frac{1}{s}\frac{\partial A_z}{\partial \phi}-\frac{\partial A_\phi}{\partial z}\right]\hat{s}+\left[\frac{\partial A_s}{\partial z}-\frac{\partial A_z}{\partial s}\right]\hat{\phi}\\ &\quad+\frac{1}{s}\left[s\frac{\partial A_\phi}{\partial s}+\frac{\partial s}{\partial s}A_\phi-\frac{\partial A_s}{\partial \phi}\right]\hat{z}\end{aligned}$$

and we can see the quantity represented by the first two terms in the z-component matches the form of Eq. 3.1.5. Rewriting, we arrive at our final answer of

$$\boxed{\begin{aligned}\vec{\nabla}\times\vec{A} &= \left[\frac{1}{s}\frac{\partial A_z}{\partial\phi}-\frac{\partial A_\phi}{\partial z}\right]\hat{s}+\left[\frac{\partial A_s}{\partial z}-\frac{\partial A_z}{\partial s}\right]\hat{\phi}\\ &\quad+\frac{1}{s}\left[\frac{\partial}{\partial s}\left(sA_\phi\right)-\frac{\partial A_s}{\partial\phi}\right]\hat{z}\end{aligned}}$$

In summary, the behavior of the del operator in **cylindrical coordinates** is given by

- *The Gradient*:

$$\vec{\nabla}f=\frac{\partial f}{\partial s}\hat{s}+\frac{1}{s}\frac{\partial f}{\partial\phi}\hat{\phi}+\frac{\partial f}{\partial z}\hat{z} \tag{3.3.2}$$

- *The Divergence*:

$$\vec{\nabla}\bullet\vec{A}=\frac{1}{s}\frac{\partial}{\partial s}\left(sA_s\right)+\frac{1}{s}\frac{\partial A_\phi}{\partial\phi}+\frac{\partial A_z}{\partial z} \tag{3.3.3}$$

- *The Curl*:

$$\begin{aligned}\vec{\nabla}\times\vec{A} &= \left[\frac{1}{s}\frac{\partial A_z}{\partial\phi}-\frac{\partial A_\phi}{\partial z}\right]\hat{s}+\left[\frac{\partial A_s}{\partial z}-\frac{\partial A_z}{\partial s}\right]\hat{\phi}\\ &\quad+\frac{1}{s}\left[\frac{\partial}{\partial s}\left(sA_\phi\right)-\frac{\partial A_s}{\partial\phi}\right]\hat{z}\end{aligned} \tag{3.3.4}$$

- *The Laplacian*:

$$\vec{\nabla}^2f=\vec{\nabla}\bullet\left(\vec{\nabla}f\right)=\frac{1}{s}\frac{\partial}{\partial s}\left(s\frac{\partial f}{\partial s}\right)+\frac{1}{s^2}\frac{\partial^2 f}{\partial\phi^2}+\frac{\partial^2 f}{\partial z^2} \tag{3.3.5}$$

Performing the above process on **spherical coordinates** results in

- *The Gradient*:

$$\vec{\nabla} f = \frac{\partial f}{\partial r}\hat{r} + \frac{1}{r}\frac{\partial f}{\partial \theta}\hat{\theta} + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}\hat{\phi} \tag{3.3.6}$$

- *The Divergence*:

$$\vec{\nabla} \bullet \vec{A} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 A_r\right) + \frac{1}{r\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\, A_\theta\right) + \frac{1}{r\sin\theta}\frac{\partial A_\phi}{\partial \phi} \tag{3.3.7}$$

- *The Curl*:

$$\begin{aligned} \vec{\nabla} \times \vec{A} \;=\; & \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial \theta}\left(\sin\theta\, A_\phi\right) - \frac{\partial A_\theta}{\partial \phi}\right]\hat{r} \\ & + \frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial A_r}{\partial \phi} - \frac{\partial}{\partial r}\left(r A_\phi\right)\right]\hat{\theta} \\ & + \frac{1}{r}\left[\frac{\partial}{\partial r}\left(r A_\theta\right) - \frac{\partial A_r}{\partial \theta}\right]\hat{\phi} \end{aligned} \tag{3.3.8}$$

- *The Laplacian*:

$$\begin{aligned} \vec{\nabla}^2 f \;=\; & \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\,\frac{\partial f}{\partial \theta}\right) \\ & + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 f}{\partial \phi^2} \end{aligned} \tag{3.3.9}$$

3.4 Arbitrary Del Operator

I could sit here and list del operations all day for every coordinate system. However, it'd be much more efficient to perform the process from Section 3.3 on an arbitrary set of coordinates. Let's say we're working in a coordinate system governed by the coordinates (q_1, q_2, q_3) with orthonormal unit vectors $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$. In general, these variables are not necessarily distance measures. We use something called a scale factor, $\{h_1, h_2, h_3\}$, to compensate. In linear algebra terms, these scale factors are the length of the non-normalized basis

vectors (i.e. the length of the basis vectors when they're not unit vectors), which have the form

$$\vec{e}_i = h_i \hat{e}_i = \frac{\partial \vec{r}}{\partial q_i}, \tag{3.4.1}$$

where $\vec{r}$ is defined by Eq. 1.1.1 and the derivative is the result of a simple coordinate transformation.

We would also like to have some idea of the form of the path (or line) element in this coordinate system. This can be easily found using the multivariable chain rule, which states

$$df = \frac{\partial f}{\partial q_1} dq_1 + \frac{\partial f}{\partial q_2} dq_2 + \frac{\partial f}{\partial q_3} dq_3$$

$$df = \sum_{i=1}^{3} \frac{\partial f}{\partial q_i} dq_i$$

$$df = \sum_{i=1}^{3} \frac{1}{h_i} \frac{\partial f}{\partial q_i} h_i dq_i$$

for some arbitrary scalar function, $f(q_1, q_2, q_3)$. If we use Eq. 2.2.2 to write this as a dot product, then

$$df = \left(\sum_{i=1}^{3} \frac{1}{h_i} \frac{\partial f}{\partial q_i} \hat{e}_i \right) \bullet \left(\sum_{i=1}^{3} h_i dq_i \hat{e}_i \right).$$

The quantity in the first set of parentheses is simply the gradient of f. Since we have included the scale factors, every term in the second set of parentheses has a unit of length making this quantity the **path element**. We can simplify the notation to get

$$df = \vec{\nabla} f \bullet d\vec{\ell} \tag{3.4.2}$$

where

$$d\vec{\ell} = h_1 dq_1 \hat{e}_1 + h_2 dq_2 \hat{e}_2 + h_3 dq_3 \hat{e}_3. \tag{3.4.3}$$

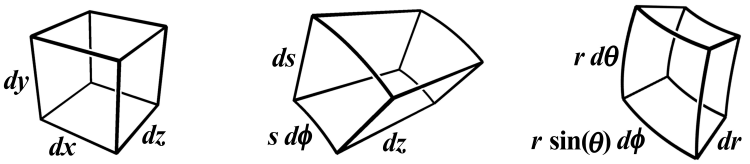


Figure 3.1: These are the volume elements of the three standard coordinate systems from Chapter 1. In order from left to right: Cartesian, Cylindrical, Spherical.

As a side note, we can integrate both sides to get

$$\int_a^b \vec{\nabla} f \bullet d\vec{\ell} = \int_a^b df = f|_{x=b} - f|_{x=a}\,, \tag{3.4.4}$$

which can be referred to from this point on as the **fundamental theorem of vector calculus** since it bears a striking resemblance to Eq. 3.1.1.

All, this talk about scale factors can be a bit confusing, so I prefer to think about them in terms of the infinitesimal volume element of the coordinate system. A volume element is made up of sides just like any other volumetric space. The volume of this element is simply the product of all three dimensions of the element (i.e. $l * w * h$) and the scale factors are the coefficients of sides. As shown in Figure 3.1, the volume element for Cartesian space is

$$dV = dx\, dy\, dz \tag{3.4.5}$$

showing scale factors of $h_x = h_y = h_z = 1$. This is why the gradient, divergence, curl, and Laplacian are all very simple. In cylindrical coordinates, however, the $d\phi$ side has a coefficient of s. This means $h_\phi = s$ and the other two scale factors are still $h_s = h_z = 1$. The cylindrical volume element is

$$dV = (ds)\,(s\, d\phi)\,(dz) = s\, ds\, d\phi\, dz. \tag{3.4.6}$$

In spherical coordinates, we find that $h_r = 1$, $h_\theta = r$, $h_\phi = r\sin\theta$, and

$$dV = (dr)\,(r\, d\theta)\,(r\sin\theta\, d\phi) = r^2 \sin\theta\, dr\, d\theta\, d\phi. \tag{3.4.7}$$

With coordinates (q_1, q_2, q_3), unit vectors $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$, and scale factors $\{h_1, h_2, h_3\}$ for an arbitrary system; the form of the del operator is given by

- *The Gradient*

$$\vec{\nabla} f = \sum_{i=1}^{3} \frac{1}{h_i} \frac{\partial f}{\partial q_i} \hat{e}_i \tag{3.4.8}$$

found easily from Eq. 3.4.2.

- *The Divergence*

$$\vec{\nabla} \bullet \vec{A} = \frac{1}{h_1 h_2 h_3} \sum_{i=1}^{3} \frac{\partial}{\partial q_i} (H_i A_i) \tag{3.4.9}$$

where $\vec{H} = (h_2 h_3)\,\hat{e}_1 + (h_3 h_1)\,\hat{e}_2 + (h_1 h_2)\,\hat{e}_3$ (the even permutations of the subscripts).

- *The Curl*

$$\vec{\nabla} \times \vec{A} = \det \begin{bmatrix} \frac{1}{h_2 h_3}\hat{e}_1 & \frac{1}{h_1 h_3}\hat{e}_2 & \frac{1}{h_1 h_2}\hat{e}_3 \\ \frac{\partial}{\partial q_1} & \frac{\partial}{\partial q_2} & \frac{\partial}{\partial q_3} \\ h_1 A_1 & h_2 A_2 & h_3 A_3 \end{bmatrix} \tag{3.4.10}$$

- *The Laplacian*

$$\vec{\nabla}^2 f = \frac{1}{h_1 h_2 h_3} \sum_{i=1}^{3} \frac{\partial}{\partial q_i} \left(H_i \frac{1}{h_i} \frac{\partial f}{\partial q_i} \right) \tag{3.4.11}$$

where $\vec{H} = (h_2 h_3)\,\hat{e}_1 + (h_3 h_1)\,\hat{e}_2 + (h_1 h_2)\,\hat{e}_3$ (the even permutations of the subscripts).

Now we have something we can apply to any coordinate system we intend on using and we don't need to look anything up. If you're not yet convinced, go ahead and try out one of the systems we've already done and see the results.

3.5 Vector Calculus Theorems

As powerful as it can be and as much insight as it can give us, the del operator may not always be the most efficient way to attack a practical problem. If this situation arises, we'll need a way to eliminate del from our equations. To do this, we'll need a slightly different perspective and a fundamental understanding of calculus.

The Divergence Theorem

Let's take another look at the divergence given in general by Eq. 3.4.9. As mentioned in Section 3.2, this is defined for a specific point in space. Theoretically, this is great because it keeps things simple, but in practice we can't really discuss specific points. All we can really do is discuss regions. To keep with the divergence, lets take this arbitrary region and divide its volume into pieces so small they might as well be points. What would these infinitesimal regions look like? Well, a volume element, of course! As we saw in Section 3.4, these volume elements look different depending on your coordinate system (some examples are given in Figure 3.1). In general, it take the form

$$dV = (h_1 dq_1)\,(h_2 dq_2)\,(h_3 dq_3) = h_1 h_2 h_3\, dq_1 dq_2 dq_3 \tag{3.5.1}$$

where $\{h_1, h_2, h_3\}$ are the scale factors and (q_1, q_2, q_3) are the coordinates.

Now let's consider the divergence throughout this volume element. From Eq. 3.4.9 and 3.5.1, we get

$$\begin{aligned} \vec{\nabla} \bullet \vec{B}\, dV &= \frac{1}{h_1 h_2 h_3} \sum_{i=1}^{3} \frac{\partial}{\partial q_i} (H_i B_i)\, h_1 h_2 h_3\, dq_1 dq_2 dq_3 \\ \vec{\nabla} \bullet \vec{B}\, dV &= \sum_{i=1}^{3} \frac{\partial}{\partial q_i} (H_i B_i)\; dq_1 dq_2 dq_3. \end{aligned} \tag{3.5.2}$$

Considering just the fist term for a moment, we have

$$\vec{\nabla} \bullet \vec{B}\, dV = \frac{\partial}{\partial q_1} (h_2 h_3 B_1)\; dq_1 dq_2 dq_3 + \ldots$$

and, if we apply the fundamental theorem of calculus (Eq. 3.1.1), we get

$$\begin{aligned} \vec{\nabla} \bullet \vec{B}\, dV &= d\,(h_2 h_3 B_1)\; dq_2 dq_3 + \ldots \\ &= (h_2 h_3 B_1)\big|_{q_1 + dq_1}\; dq_2 dq_3 - (h_2 h_3 B_1)\big|_{q_1}\; dq_2 dq_3 + \ldots \end{aligned}$$

If we regroup some of the quantities, this results in

$$\vec{\nabla} \bullet \vec{B}\, dV = B_1\big|_{q_1 + dq_1}\, (h_2 h_3\, dq_2 dq_3)\big|_{q_1 + dq_1} - B_1\big|_{q_1}\, (h_2 h_3\, dq_2 dq_3)\big|_{q_1} + \ldots$$

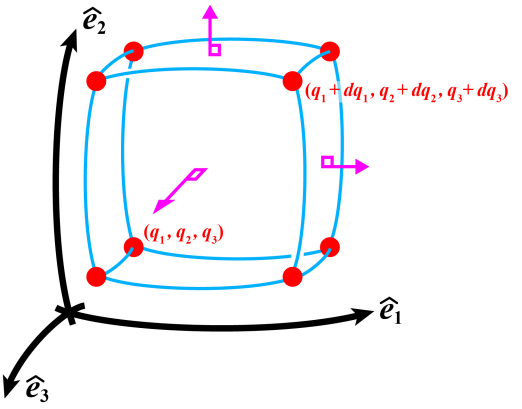


Figure 3.2: This is a representation of an arbitrary volume element. The orthogonal vectors for each of the surfaces facing the reader are also shown. The back bottom left corner is labeled (q_1, q_2, q_3) and the front top right corner is labeled $(q_1 + dq_1, q_2 + dq_2, q_3 + dq_3)$ to show that its volume matches that given by Eq. 3.5.1.

Taking a look at Figure 3.2, we can see the first of these two terms corresponds to the right surface of the volume element located at $q_1 + dq_1$ and the second term corresponds to the left surface located at q_1. Each of these surfaces spans an area of $dA_1 = (h_2 dq_2)\,(h_3 dq_3) = h_2 h_3 dq_2 dq_3$ evaluated at their location along q_1. This simplifies the above relationship to

$$\vec{\nabla} \bullet \vec{B}\, dV = B_1|_{q_1+dq_1}\ dA_1|_{q_1+dq_1} - B_1|_{q_1}\ dA_1|_{q_1} + \ldots$$

Any area is represented by a vector orthogonal to its surface (i.e. $d\vec{A} = \hat{n} \cdot dA$). In the case of dA_1, this orthogonal vector is $d\vec{A}_1 = \hat{e}_1 \cdot dA_1$. If the area element is a vector, the above looks a lot like the definition of the dot product (Eq. 2.2.2). Taking advantage of this, we get

$$\vec{\nabla} \bullet \vec{B}\, dV = \vec{B} \bullet d\vec{A}_1\Big|_{q_1+dq_1} + \vec{B} \bullet d\vec{A}_1\Big|_{q_1} + \ldots \tag{3.5.3}$$

where we've lost our three negative signs because the angle between $\vec{B}$ and $d\vec{A}$ for those surfaces is 180° (because $d\vec{A}$ always points outward from the volume enclosed). The cosine in Eq. 2.2.1 takes care of the sign for us.

Originally, in Eq. 3.5.2, we had three terms. Now, with Eq. 3.5.3, we have six terms each corresponding to a different surface of the volume element (which is composed of six surfaces). Since this process has occurred for all six

terms and these six terms together completely enclose the volume element, we can rewrite Eq. 3.5.3 as

$$\vec{\nabla} \bullet \vec{B}\, dV = \oint_{dV} \vec{B} \bullet d\vec{A}. \tag{3.5.4}$$

If we're going to make this practical, it should apply to the entire region, not just the volume element. To do this, we simply add up (with an integral) all the elements that compose the region. But what happens to the right side of Eq. 3.5.4? If the region is composed of volume elements, then those elements are all touching such that they completely fill the region. For the surface (area) elements in contact with other surface elements within the region, their $\vec{B} \bullet d\vec{A}$'s will all cancel because all of their $d\vec{A}$'s will be exactly opposite. This means only the surface elements *not* in contact with other surface elements will add to the integral on the right in Eq. 3.5.4. These surface elements are simply the ones on the outside of the region (i.e. we only need to integrate over the outside surface of the region). Therefore, Eq. 3.5.4 becomes

$$\int \vec{\nabla} \bullet \vec{B}\, dV = \oint_{V} \vec{B} \bullet d\vec{A}\,. \tag{3.5.5}$$

We call this the **Divergence Theorem** and it is true for any arbitrary region V enclosed by a surface A.

You may be asking yourself why we didn't just start with the volume of the entire region from the beginning. Why did we do all this stuff with the volume element instead? The answer is simple: We know what the volume element looks like. We know it has six faces and that these faces have a very definite size and shape within the coordinate system. The same cannot be said about the entire region because it's completely arbitrary. When we say arbitrary, we don't just mean that the system we apply this to could have any configuration. We mean that, even with a particular system, we can really choose a region with any shape, size, orientation, or location we wish and Eq. 3.5.5 still applies.

The Curl Theorem

Let's take another look at the curl given in general by Eq. 3.4.10. As mentioned in Section 3.2, this is defined for a specific point in space just like

the divergence. However, in practical situations, exact points are difficult to discuss. When it came to the divergence, it was regions of volume we really discuss. With the curl, it's areas of circulation. Again, keeping with the idea of a single point, let's divide our area into pieces so small that they might as well be points. These pieces would correspond to the surface elements which look different depending on your coordinate system (some examples correspond to the faces of the volume elements in Figure 3.1).

We'll want things to be as general as possible, so we'll still use the coordinates (q_1, q_2, q_3). However, to keep things simple, we'll choose a particular surface element from Figure Figure 3.2 given by

$$d\vec{A}_3 = (h_1 dq_1)\,(h_2 dq_2)\,\hat{e}_3 = h_1 h_2\, dq_1 dq_2\,\hat{e}_3 \tag{3.5.6}$$

where $\{h_1, h_2, h_3\}$ are the scale factors and $\hat{e}_3$ is the vector orthogonal to the surface element. Now we'll consider the curl on that surface element given by

$$\begin{aligned}
\left(\vec{\nabla}\times\vec{B}\right)\bullet d\vec{A}_3 &= \left(\vec{\nabla}\times\vec{B}\right)_3 dA_3 \\
&= \frac{1}{h_1 h_2}\left[\frac{\partial}{\partial q_1}(h_2 B_2) - \frac{\partial}{\partial q_2}(h_1 B_1)\right] h_1 h_2\, dq_1 dq_2 \\
&= \left[\frac{\partial}{\partial q_1}(h_2 B_2) - \frac{\partial}{\partial q_2}(h_1 B_1)\right] dq_1 dq_2 \\
&= \frac{\partial}{\partial q_1}(h_2 B_2)\; dq_1 dq_2 - \frac{\partial}{\partial q_2}(h_1 B_1)\; dq_1 dq_2
\end{aligned}$$

where we have applied Eqs. 3.4.10 and 3.5.6. If we apply the fundamental theorem of calculus (Eq. 3.1.1), we get

$$\begin{aligned}
\left(\vec{\nabla}\times\vec{B}\right)\bullet d\vec{A}_3 &= d(h_2 B_2)\; dq_2 - d(h_1 B_1)\; dq_1 \\
&= (h_2 B_2)\big|_{q_1+dq_1}\, dq_2 - (h_2 B_2)\big|_{q_1}\, dq_2 \\
&\quad - (h_1 B_1)\big|_{q_2+dq_2}\, dq_1 + (h_1 B_1)\big|_{q_2}\, dq_1.
\end{aligned}$$

If we regroup some of the quantities, this results in

$$\begin{aligned}
\left(\vec{\nabla}\times\vec{B}\right)\bullet d\vec{A}_3 &= B_2\big|_{q_1+dq_1}\,(h_2 dq_2)\big|_{q_1+dq_1} - B_2\big|_{q_1}\,(h_2 dq_2)\big|_{q_1} \\
&\quad - B_1\big|_{q_2+dq_2}\,(h_1 dq_1)\big|_{q_2+dq_2} + B_1\big|_{q_2}\,(h_1 dq_1)\big|_{q_2}.
\end{aligned}$$

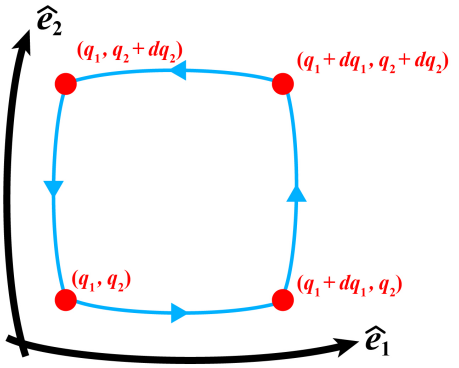


Figure 3.3: This is a representation of the arbitrary surface element with orthogonal vector of $\hat{e}_3$. The corner point between which we integrate are labeled with coordinates given in form (q_1, q_2) and it is assumed that all points in this diagram have the same q_3 value.

Taking a look at Figure 3.3, we can see the first term corresponds to the right part of the curve bounding the surface element. Furthermore, the second term corresponds to the left part, the third term to the top part, and the fourth term to the bottom part. This means the entire curve enclosing the surface element is represented. Just as with the divergence theorem, we see the terms match the form of the dot product given by Eq. 2.2.2. Since the direction we assign to this curve is completely arbitrary, let's keep things consistent with the right-hand rule and choose counterclockwise. This way the negative signs in the second and third terms are explained by the direction of the curve being opposite from the first and fourth terms, respectively (we're defining up and to the right as positive). All this considered and defining $d\ell_i = h_i\, dq_i$, we can rewrite as

$$\begin{aligned}\left(\vec{\nabla}\times\vec{B}\right)\bullet d\vec{A}_3 &= \left(\vec{B}\bullet d\vec{\ell}_2\right)\Big|_{q_1+dq_1} + \left(\vec{B}\bullet d\vec{\ell}_2\right)\Big|_{q_1} \\ &\quad + \left(\vec{B}\bullet d\vec{\ell}_1\right)\Big|_{q_2+dq_2} + \left(\vec{B}\bullet d\vec{\ell}_1\right)\Big|_{q_2}. \\ \left(\vec{\nabla}\times\vec{B}\right)\bullet d\vec{A}_3 &= \oint_{dA_3}\vec{B}\bullet d\vec{\ell}. \end{aligned} \tag{3.5.7}$$

The result in Eq. 3.5.7 is only true of areas constructed of surface elements with orthogonal vector $\hat{e}_3$. However, nothing was really special about this particular surface element. We could have just as easily (and in exactly the

same way) found the curl on one of the other elements given by

$$d\vec{A}_1 = (h_2 dq_2)\,(h_3 dq_3)\,\hat{e}_1 = h_2 h_3\, dq_2 dq_3\,\hat{e}_1 \tag{3.5.8}$$

or

$$d\vec{A}_2 = (h_1 dq_1)\,(h_3 dq_3)\,\hat{e}_2 = h_1 h_3\, dq_1 dq_3\,\hat{e}_2. \tag{3.5.9}$$

This would have resulted in

$$\left(\vec{\nabla}\times\vec{B}\right)\bullet d\vec{A}_1 = \oint_{dA_1}\vec{B}\bullet d\vec{\ell} \tag{3.5.10}$$

or

$$\left(\vec{\nabla}\times\vec{B}\right)\bullet d\vec{A}_2 = \oint_{dA_2}\vec{B}\bullet d\vec{\ell}, \tag{3.5.11}$$

respectively.

Eqs. 3.5.7, 3.5.10, and 3.5.11 describe the three possible orthogonal orientations provided by our three-dimensional space. This means any surface can be constructed of some combination of these surface elements (or projections onto these elements). This includes the practical area with which we started our discussion. To do this, we simply add up (with an integral) all the elements that compose the area. But what happens to the right side of the equation? If the area is composed of surface elements, then those elements are all touching such that they completely fill the area. Many of the curve elements that enclose each surface element are in contact with curve elements of other surfaces elements For those curve elements, their $\vec{B}\bullet d\vec{\ell}$'s will all cancel because all of their $d\vec{\ell}$'s will be exactly opposite. This means only the curve elements *not* in contact with other curve elements will add to the integral on the right. These curve elements are simply the ones on the outside of the area (i.e. we only need to integrate over the outside curve that encloses the area). Therefore, our general equation becomes

$$\int\left(\vec{\nabla}\times\vec{B}\right)\bullet d\vec{A} = \oint_{A}\vec{B}\bullet d\vec{\ell}. \tag{3.5.12}$$

We call this the **Curl Theorem** (or often **Stokes Theorem**) and it is true for any arbitrary area A enclosed by a curve ℓ.

You may be asking yourself why we didn't just start with the area of the entire region from the beginning. Why did we do all this stuff with the surface elements instead? The answer is simple: We know what the surface elements look like. We know it has four sides and that these sides have a very definite size and shape within the coordinate system. The same cannot be said about the entire area because it's completely arbitrary. When we say arbitrary, we don't just mean that the system we apply this to could have any configuration. We mean that, even with a particular system, we can really choose an area with any shape, size, orientation, or location we wish and Eq. 3.5.12 still applies.

Chapter 4

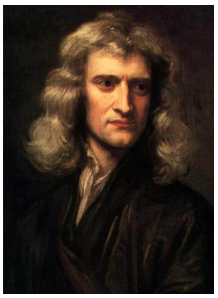
Lagrangian Mechanics

4.1 A Little History...

Classical mechanics was given birth with the publication of *Philosophiæ Naturalis Principia Mathematica* (Latin for "*Mathematical Principles of Natural Philosophy*") by Sir Isaac Newton in 1687. It finally laid to rest Aristotle's view of motion and was a basic framework for the physics to come over the following century. The Principia contained Newton's universal law of gravitation as well as Newton's three laws of motion. Together, they connect the Earth with the Heavens in one construction.

The only disadvantage to Newton's laws is they are written in terms of vector quantities, quantities which depend on direction. This makes the mathematics behind them a bit of a hassle at times and arguably less elegant. A couple years after the publication of the Principia, Gottfried Wilhelm von Leibniz (the German mathematician that invented calculus independently from Newton) began to voice opinions of a scalar quantity he had noticed which he called **vis viva** (Latin for "force of life"). This scalar would eventually become known as **kinetic energy**. The idea of scalar quantities was opposed by Newton for quite some time because he felt it was inconsistent with his conservation of momentum.

In 1788, Joseph Louis Lagrange published *Mécanique Analytique* (Latin for "*Analytical Mechanics*") where he derived his equations. These equations were contrasted from Newton's because they were formulated entirely in terms of scalar quantities. However, the term **energy** was not used to describe them until 1807 by Thomas Young and the conservation of energy


Isaac Newton


Gottfried Leibniz

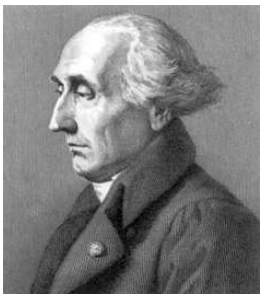

Joseph Louis Lagrange

Figure 4.1: These people were important in the development leading up to Lagrange's equation.

was not formally written until 1847 by Hermann von Helmholtz. This would suggest Lagrange didn't have much background as to the nature of these scalar quantities, but we know from his own words that he didn't mind.

> "No diagrams will be found in this work. The methods that I explain in it require neither constructions nor geometrical or mechanical arguments, but only the algebraic operations inherent to a regular and uniform process. Those who love Analysis will, with joy, see mechanics become a new branch of it and will be grateful to me for having extended this field."

In Section 4.2, a derivation is presented using our modern understanding of these quantities. The intent is to present it in a similar fashion to Lagrange, yet a little less abstract than I expect Lagrange's presentation was.

4.2 Derivation of Lagrange's Equation

Deriving the highly useful Lagrange's equation requires little more than Newton's second law and the definition of work. We'll simplify the derivation by assuming the system is composed of only one body of mass, m. Later on, we'll see how this derivation can be easily generalized to describe a multiple body system. The definition of work is

$$dW \equiv \vec{F} \bullet d\vec{r} \tag{4.2.1}$$

where $\vec{r}$ is the position vector of m and $\vec{F}$ is the force on m. However, in general, force is a function of space and time. This detail can complicate the

derivation, so to make it easier we'll consider only the spacial components of the position $\vec{r}$ by setting the change in time to be exactly zero. No, we don't mean infinitesimally small, we mean zero. Under these non-realistic or *virtual* conditions, $d\vec{r}$ becomes $\delta\vec{r}$ or the virtual displacement (because m doesn't really displace in a zero time interval). Even though all of this is pretend, we can still get some very useful results if we can make the virtual quantities drop out later on in the derivation. Therefore, we have the definition of virtual work

$$\delta W = \vec{F} \bullet \delta\vec{r}. \tag{4.2.2}$$

If our system is free of **non-conservative forces**, then we can write force in terms of the **potential energy**, V, as

$$\vec{F} = -\vec{\nabla}V \tag{4.2.3}$$

where $\vec{\nabla}$ is the del operator (defined in Chapter 3). In this particular case, it is called the gradient which measures the change in the scalar quantity V through space (i.e. it is a vector derivative with respect to space). With this substitution, virtual work becomes

$$\delta W = -\vec{\nabla}V \bullet \delta\vec{r}. \tag{4.2.4}$$

Part of the beauty of Lagrange's equation is that it works with a set of **generalized coordinates** rather than the three dimensions represented by $\vec{r}$. Generalized coordinates, q_i, are a set of coordinates that are natural to the system and are not necessarily limited to three, which becomes clear in more complex examples. For this reason, these generalized coordinates are not referred to as dimensions but **degrees of freedom** because they represent the amount of freedom our system has to move. If we write $d\vec{r}$ in terms of q_i using a coordinate transformation, then Eq. 4.2.4 becomes

$$\delta W = -\vec{\nabla}V \bullet \sum_{i=1}^{n} \frac{\partial\vec{r}}{\partial q_i}\delta q_i$$

where n is the number of degrees of freedom of the system (i.e. the number of generalized coordinates).

The dot product is simply a sum of the products of the vector components (defined by Eq. 2.2.2) and the components of the gradient are defined by Eq.

3.2.2 (the standard is to start with Cartesian coordinates). Therefore, work becomes

$$\delta W = -\sum_{j=1}^{3}\sum_{i=1}^{n}\frac{\partial V}{\partial r_j}\frac{\partial r_j}{\partial q_i}\delta q_i.$$

The new summation only has three terms because $\vec{r}$ is a position vector in 3-space. We can now cancel out our original coordinate system leaving us with

$$\delta W = -\sum_{i=1}^{n}\frac{\partial V}{\partial q_i}\delta q_i \tag{4.2.5}$$

only in terms of the generalized coordinates.

We can also make a substitution in Eq. 4.2.2 using Newton's second law,

$$\vec{F} = \dot{\vec{p}} = m\ddot{\vec{r}}, \tag{4.2.6}$$

and we get

$$\delta W = m\ddot{\vec{r}} \bullet \delta\vec{r}.$$

Again, we can write the dot product as a summation and work becomes

$$\delta W = m\sum_{j=1}^{3}\ddot{r}_j\delta r_j. \tag{4.2.7}$$

Also as before, we use a coordinate transformation to write work as

$$\delta W = m\sum_{i=1}^{n}\sum_{j=1}^{3}\ddot{r}_j\frac{\partial r_j}{\partial q_i}\delta q_i$$

We can now take advantage of the product rule (defined by Eq. 3.1.5) and that time derivatives commute with spatial derivatives

$$\frac{d}{dt}\left(\dot{r}_j\frac{\partial r_j}{\partial q_i}\right) = \ddot{r}_j\frac{\partial r_j}{\partial q_i} + \dot{r}_j\frac{\partial \dot{r}_j}{\partial q_i}$$

$$\Rightarrow \ddot{r}_j\frac{\partial r_j}{\partial q_i} = \frac{d}{dt}\left(\dot{r}_j\frac{\partial r_j}{\partial q_i}\right) - \dot{r}_j\frac{\partial \dot{r}_j}{\partial q_i}$$

and work becomes

$$\delta W = m \sum_{i=1}^{n} \sum_{j=1}^{3} \left[\frac{d}{dt} \left(\dot{r}_j \frac{\partial r_j}{\partial q_i} \right) - \dot{r}_j \frac{\partial \dot{r}_j}{\partial q_i} \right] \delta q_i.$$

Again, time derivatives commute with spatial derivatives. Therefore, we can perform the operation

$$\frac{\partial r_j}{\partial q_i} = \frac{\frac{d}{dt}\left(\partial r_j\right)}{\frac{d}{dt}\left(\partial q_i\right)} = \frac{\partial \dot{r}_j}{\partial \dot{q}_i},$$

which can be used as a substitution in the above relationship for work. We now get

$$\delta W = m \sum_{i=1}^{n} \sum_{j=1}^{3} \left[\frac{d}{dt} \left(\dot{r}_j \frac{\partial \dot{r}_j}{\partial \dot{q}_i} \right) - \dot{r}_j \frac{\partial \dot{r}_j}{\partial q_i} \right] \delta q_i.$$

We can use the derivative chain rule (Eq. 3.1.2)

$$\frac{d}{dx}\left(u^2\right) = \frac{d}{du}\left(u^2\right)\frac{du}{dx} = 2u\frac{du}{dx} \quad \Rightarrow u\frac{du}{dx} = \frac{d}{dx}\left(\frac{1}{2}u^2\right) \tag{4.2.8}$$

to change the variable with which we're differentiating and work becomes

$$\delta W = m \sum_{i=1}^{n} \sum_{j=1}^{3} \left[\frac{d}{dt}\frac{\partial}{\partial \dot{q}_i} \left(\frac{1}{2}\dot{r}_j^2 \right) - \frac{\partial}{\partial q_i} \left(\frac{1}{2}\dot{r}_j^2 \right) \right] \delta q_i.$$

Bringing the m and the summation over the index j inside the derivatives, we get

$$\delta W = \sum_{i=1}^{n} \left[\frac{d}{dt}\frac{\partial}{\partial \dot{q}_i} \left(\sum_{j=1}^{3} \frac{1}{2} m \dot{r}_j^2 \right) - \frac{\partial}{\partial q_i} \left(\sum_{j=1}^{3} \frac{1}{2} m \dot{r}_j^2 \right) \right] \delta q_i. \tag{4.2.9}$$

The summation over j is now simply the kinetic energy, K, of the system. Applying this definition to Eq. 4.2.9, we get

$$\delta W = \sum_{i=1}^{n} \left[\frac{d}{dt} \left(\frac{\partial K}{\partial \dot{q}_i} \right) - \frac{\partial K}{\partial q_i} \right] \delta q_i. \tag{4.2.10}$$

Now we can bring Eqs. 4.2.5 and 4.2.10 together and we get

$$-\sum_{i=1}^{n}\frac{\partial V}{\partial q_i}\delta q_i = \sum_{i=1}^{n}\left[\frac{d}{dt}\left(\frac{\partial K}{\partial \dot{q}_i}\right) - \frac{\partial K}{\partial q_i}\right]\delta q_i$$

$$\Rightarrow \sum_{i=1}^{n}\left[\frac{\partial\left(K-V\right)}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial K}{\partial \dot{q}_i}\right)\right]\delta q_i = 0. \tag{4.2.11}$$

If the potential energy, V, is only a function of position (which it is by definition), then we know

$$\frac{\partial V}{\partial \dot{q}_i} = 0. \tag{4.2.12}$$

This allows us to do something I like to call *voodoo math* (with a little foresight; we can add zeros, multiply by ones, add and subtract constants, etc. to simplify a mathematical expression) and Eq. 4.2.11 becomes

$$\sum_{i=1}^{n}\left[\frac{\partial\left(K-V\right)}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial\left(K-V\right)}{\partial \dot{q}_i}\right)\right]\delta q_i = 0.$$

Since this mathematical statement must be true for all systems of general coordinates, we have

$$\frac{\partial\left(K-V\right)}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial\left(K-V\right)}{\partial \dot{q}_i}\right) = 0. \tag{4.2.13}$$

Note that the virtual displacements have disappeared from our equation, which is exactly what we needed to happen so this could all make sense.

Eq. 4.2.13 is called **Lagrange's equation**, but we can do better. Let's define a **Lagrangian** as $\mathcal{L} = KE - PE = K - V$ so that Eq. 4.2.13 can be written simply as

$$\frac{\partial \mathcal{L}}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right) = 0 \tag{4.2.14}$$

where q_i are the generalized coordinates and $\dot{q}_i$ are the generalized velocities. The index i indicates there are as many of these equations for your system as you have generalized coordinates, so you will always have as many equations as unknowns (i.e. a solvable system). If the generalized coordinate is a linear distance measure, then Eq. 4.2.14 results in force terms. If the generalized coordinate is an angle measure, then Eq. 4.2.14 results in torque terms. The solutions are *always* the equations of motion for a given system.

4.3 Generalizing for Multiple Bodies

As mentioned before, this derivation can be easily generalized to a system of N bodies to arrive at exactly the result given by Eq. 4.2.14. Let's designate the force on the k^{th} body as

$$\vec{F}_k = \dot{\vec{p}}_k = m_k \ddot{\vec{r}}_k,$$

which is similar to Eq. 4.2.6. Therefore, the virtual work on the k^{th} body is

$$\delta W_k = m_k \ddot{\vec{r}}_k \bullet \delta \vec{r}_k$$

and the total virtual work of the system is the sum of these terms given by

$$\delta W = \sum_{k=1}^{N} m_k \ddot{\vec{r}}_k \bullet \delta \vec{r}_k.$$

We can write this similar to Eq. 4.2.7 resulting in

$$\delta W = \sum_{k=1}^{N} m_k \sum_{j=1}^{3} \ddot{r}_{kj} \delta r_{kj}.$$

and, after a coordinate transformation to generalized coordinates, it becomes

$$\delta W = \sum_{k=1}^{N} m_k \sum_{i=1}^{n} \sum_{j=1}^{3} \ddot{r}_{kj} \frac{\partial r_{kj}}{\partial q_i} \delta q_i. \tag{4.3.1}$$

We do not need to index the generalized coordinates with k because we are already keeping a separate set of them for each body. A one-body system has, at most, 3 degrees of freedom. A two-body system has, at most, 6 degrees of freedom. Therefore, an N-body system has, at most, $3N$ degrees of freedom. As mentioned in Section 4.2, we are not limited to three independent variables.

We can begin to see why this would have gotten rather complicated. Cluttering our derivation with lots of summations and indices would have certainly been complete, but we may have missed the beauty with such rigor. Based on the process given in Section 4.2, we can see the extra summation in Eq. 4.3.1 and index will not affect the steps. Eq. 4.2.9 will appear as

$$\delta W = \sum_{i=1}^{n} \left[\frac{d}{dt} \frac{\partial}{\partial \dot{q}_i} \left(\sum_{k=1}^{N} \sum_{j=1}^{3} \frac{1}{2} m_k \dot{r}_{kj}^2 \right) - \frac{\partial}{\partial q_i} \left(\sum_{k=1}^{N} \sum_{j=1}^{3} \frac{1}{2} m_k \dot{r}_{kj}^2 \right) \right] \delta q_i.$$

The parenthetical quantity is simply the kinetic energy of the whole system and can still be defined as K. Under this definition, our new virtual work becomes exactly Eq. 4.2.10 and still ultimately results in Eq. 4.2.14 given that we define $\mathcal{L}$ as the Lagrangian of the whole system of N bodies.

4.4 Applications of Lagrange's Equation

There is a methodical process to solving problems using Lagrange's equations:

1. *Determine the best set of generalized coordinates for the system.* There are an infinite number of these sets, but we can make things easier by making a good choice. The best choice will have the minimum number of degrees of freedom for the system.

2. *Write out the coordinate transformations.* In other words, write the Cartesian coordinates of each object in terms of the generalized coordinates and take each of their first time-derivatives.

3. *Use the coordinate transformations to write out the potential and kinetic energy of the system in terms the generalized coordinates.* If you have multiple bodies in the system, then you can find the total by adding the corresponding energy from all the bodies together.

4. *Find the Lagrangian of the system.* Recall $\mathcal{L} = K - V$.

5. *Plug the Lagrangian into Lagrange's equation.* See Eq. 4.2.14.

Example 4.4.1

A solid ball (mass m and radius R), starting from rest, rolls without slipping down an platform inclined at an angle ϕ from the floor.

1. We can define x as the distance the ball has traveled down the incline and θ as the angle through which the ball has rotated. This would constitute a set of generalized coordinates. However, the ball has a constraint that it doesn't slip on the surface of the incline. Therefore, x and θ are related by $x = R\theta$, an **equation of constraint.** This means only one of them is required. We'll choose x.

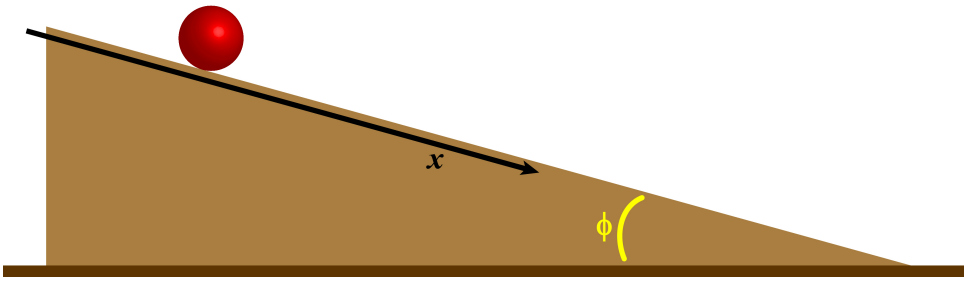


Figure 4.2: The ball in this figure is rolling without slipping down the platform. The displacement from the top, x, is labeled as well as the angle of inclination, ϕ, of the platform.

2. Since we only have one object with one generalized coordinate, $x = x$ is the coordinate transformation. It seems kinda trivial, doesn't it? Rest assured they will be more interesting in more complex examples.

3. The potential and kinetic energy of the ball are given by

$$V = mgh = -mgx\sin\phi$$

and, since $I = \frac{2}{5}mR^2$ for a solid sphere and $x = R\theta \Rightarrow \dot{x} = R\dot{\theta}$, we get

$$K = \tfrac{1}{2}mv^2 + \tfrac{1}{2}I\omega^2 = \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{2}I\dot{\theta}^2 = \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{5}m\dot{x}^2 = \tfrac{7}{10}m\dot{x}^2.$$

4. The Lagrangian is

$$\mathcal{L} = K - V = \tfrac{7}{10}m\dot{x}^2 - (-mgx\sin\phi) = \tfrac{7}{10}m\dot{x}^2 + mgx\sin\phi.$$

5. Plugging this into Lagrange's equation, we get

$$\frac{\partial\mathcal{L}}{\partial x} - \frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{x}}\right) = 0$$

$$mg\sin\phi - \tfrac{d}{dt}\left(\tfrac{7}{5}m\dot{x}\right) = mg\sin\phi - \tfrac{7}{5}m\ddot{x} = 0$$

$$\ddot{x} = \frac{5}{7}g\sin\phi.$$

This is the acceleration of the ball as it travels down the incline. Under normal circumstances we would integrate this twice to find the function $x(t)$, but because the acceleration is constant we already know this will result in

$$\begin{aligned} x(t) &= \tfrac{1}{2}a_x t^2 + v_{0x}t + x_0 = \tfrac{1}{2}\ddot{x}t^2 + \dot{x}(0)\,t + x(0) \\ &= \tfrac{1}{2}\ddot{x}t^2 = \tfrac{1}{2}\left(\tfrac{5}{7}g\sin\phi\right)t^2 \end{aligned}$$

$$\boxed{x(t) = \left(\frac{5}{14}g\sin\phi\right)t^2}.$$

If you want to know how the ball is rotating at a given time, then

$$\theta(t) = \frac{x(t)}{R} = \left(\frac{5g\sin\phi}{14R}\right)t^2.$$

This is exactly the result you would get via Newton's laws.

Example 4.4.2

An object with a mass m is moving within the gravitational influence of the sun ($M_\odot = 1.99 \times 10^{30}$ kg) such that $m \ll M_\odot$.

1. The position of m is represented by (r, θ) in cylindrical coordinates. Neither of these coordinates is necessarily constant with the information provided. If there is any motion at all, then θ is changing. The value of r is only constant for a circular orbit and, as close as some of the planets may get to this, most orbits are not circular. Therefore, our generalized coordinates, q_i, are (r, θ).

2. Based on Figure 4.3, we can write the coordinate transformations as

$$\left\{\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned}\right\}$$

and the first time-derivatives are

$$\left\{\begin{aligned} \dot{x} &= \dot{r}\cos\theta - r\dot{\theta}\sin\theta \\ \dot{y} &= \dot{r}\sin\theta + r\dot{\theta}\cos\theta \end{aligned}\right\}.$$

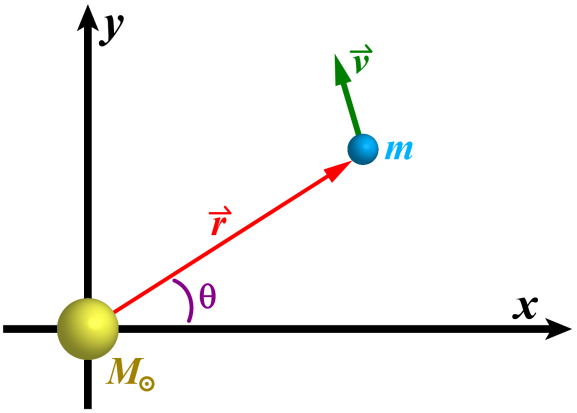


Figure 4.3: The sun has been placed at the origin in the coordinate system for convenience. The position, $\vec{r}$, is arbitrary and the velocity, $\vec{v}$, is shown for that position. Note that $\vec{v}$ is not perpendicular to $\vec{r}$ since this only true for a circular path.

3. The potential energy possessed by the object is due to the gravitational potential created by the sun and is given by

$$V = -G\frac{M_\odot m}{r}$$

and kinetic energy is given by

$$\begin{aligned} K &= \frac{1}{2}mv^2 = \frac{1}{2}m\left(\dot{x}^2 + \dot{y}^2\right) \\ &= \frac{1}{2}m\left[\left(\dot{r}\cos\theta - r\dot{\theta}\sin\theta\right)^2 + \left(\dot{r}\sin\theta + r\dot{\theta}\cos\theta\right)^2\right] \\ &= \frac{1}{2}m\Big(\dot{r}^2\cos^2\theta - r\dot{r}\dot{\theta}\sin\theta\cos\theta + r^2\dot{\theta}^2\sin^2\theta \\ &\quad + \dot{r}^2\sin^2\theta + r\dot{r}\dot{\theta}\sin\theta\cos\theta + r^2\dot{\theta}^2\cos^2\theta\Big) \end{aligned}$$

and, since $\sin^2\theta + \cos^2\theta = 1$,

$$K = \frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right).$$

4. The Lagrangian is

$$\begin{aligned}\mathcal{L} &= K - V \\ &= \frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right) - \left[-G\frac{M_\odot m}{r}\right] \\ &= \frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right) + G\frac{M_\odot m}{r}.\end{aligned}$$

5. The Lagrange's equations applied to this example are

$$\left\{\begin{aligned}\frac{\partial\mathcal{L}}{\partial r} - \frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{r}}\right) &= 0 \\ \frac{\partial\mathcal{L}}{\partial\theta} - \frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{\theta}}\right) &= 0\end{aligned}\right\}$$

$$\left\{\begin{aligned}mr\dot{\theta}^2 - G\frac{M_\odot m}{r^2} - \frac{d}{dt}\left(m\dot{r}\right) &= 0 \\ 0 - \frac{d}{dt}\left(-mr^2\dot{\theta}\right) &= 0\end{aligned}\right\}$$

$$\left\{\begin{aligned}mr\dot{\theta}^2 - G\frac{M_\odot m}{r^2} - m\ddot{r} &= 0 \\ \frac{d}{dt}\left(-mr^2\dot{\theta}\right) &= 0\end{aligned}\right\}.$$

$$\left\{\begin{aligned}\ddot{r} - r\dot{\theta}^2 + \frac{GM_\odot}{r^2} &= 0 \\ \frac{d}{dt}\left(r^2\dot{\theta}\right) &= 0\end{aligned}\right\}.$$

Let's take a look at the second equation of motion. It implies that $r^2\dot{\theta}$ is constant. This result is important for two reason. First, we know based on the derivation of Eq. 4.2.14 that the second term represent the change in momentum of the system. In this case, $L = mr^2\dot{\theta}$ (or $\ell = r^2\dot{\theta}$) represents the angular momentum of the system because we differentiated by an angle

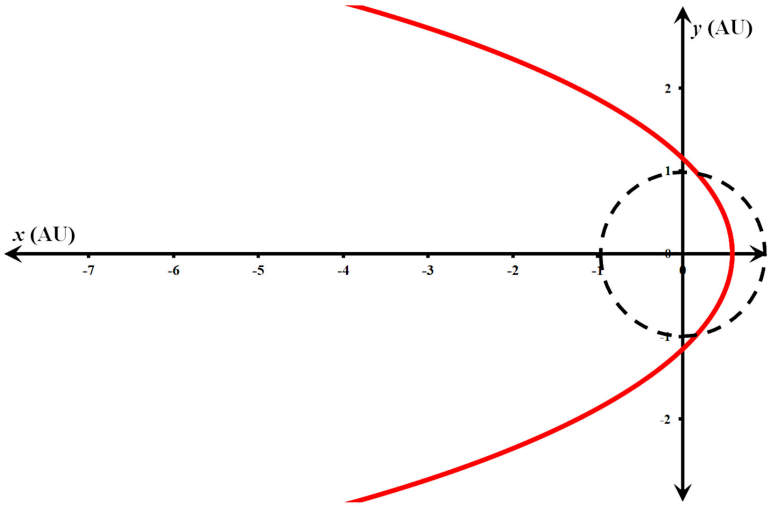


Figure 4.4: This is the elliptical orbit of Halley's comet. It has been scaled to make visible the entire orbit of Earth at 1 AU. The sun still does not have visible size at this scale. Note: this diagram does not indicate orientation.

rather than a distance. Therefore, angular momentum is conserved by a central force (e.g. gravity or electrostatics). Second, the area swept out by the object in its orbit is given by

$$\begin{aligned} dA &= \frac{1}{2}r^2 d\theta \\ \frac{dA}{dt} &= \frac{1}{2}r^2\frac{d\theta}{dt} = \frac{1}{2}r^2\dot{\theta} = \frac{1}{2}\ell = \text{constant}. \end{aligned} \tag{4.4.1}$$

This is Kepler's second law of planetary motion. If these equations are taking us in this direction, let's find out where the other one leads.

The first equation of motion can be simplified given that $\ell = r^2\dot{\theta}$ resulting in

$$\begin{aligned} \ddot{r} - r\left(\frac{\ell}{r^2}\right)^2 + \frac{GM_\odot}{r^2} &= 0 \\ \ddot{r} - \frac{\ell^2}{r^3} + \frac{GM_\odot}{r^2} &= 0. \end{aligned}$$

At first glance, this differential equation might seem challenging. However, with a very simple change of variable given by $r = u^{-1}$, it will become

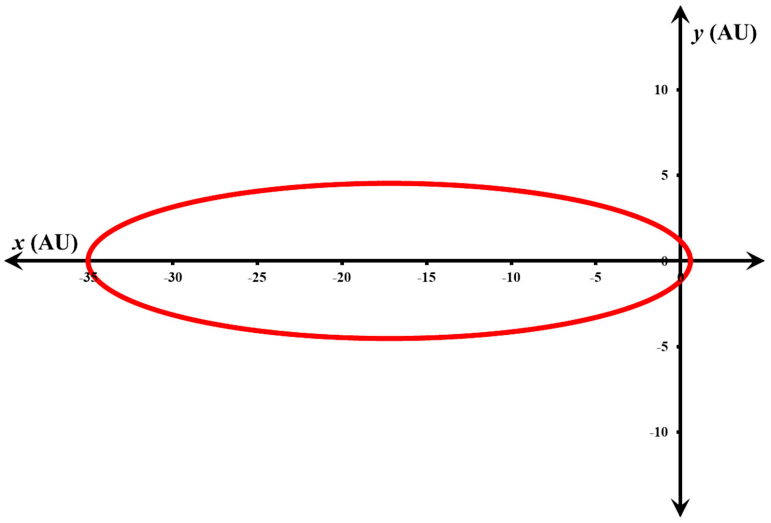


Figure 4.5: This is the elliptical orbit of Halley's comet. It has been scaled to make the comet's entire orbit visible.

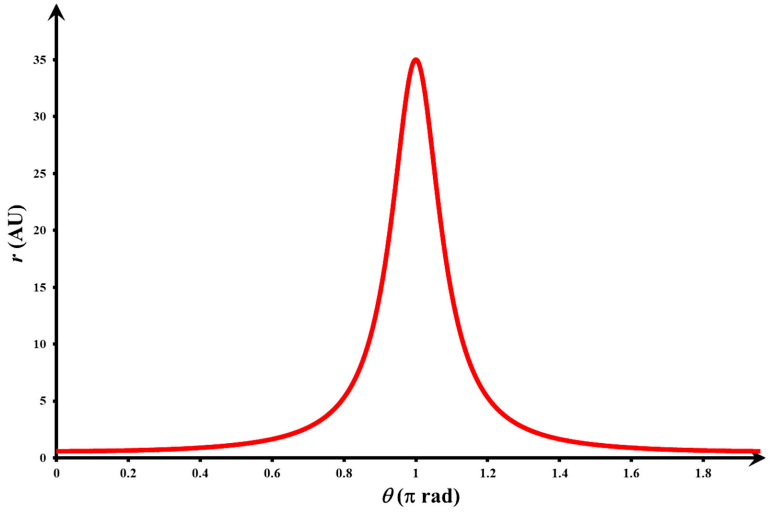


Figure 4.6: This is a graph of the radius, r, (distance from the sun) as a function of orbital angle, θ. It also indicates that the radius has a higher rate of change at π rad (i.e. the aphelion).

a very straight forward equation. If you don't have the knack for solving differential equations yet, don't worry. Solving them really only comes down to two things: making good guesses and knowing where you're going. These will come to you with experience. Well, now that we've got our good guess out of the way, where are we going? We would like r to be function of θ rather than t because that will help us get an idea of the shape of the object's path (maybe an ellipse?). Using $\dot{\theta} = \ell/r^2 = \ell u^2$ and the chain rule, we get a first derivative substitution of

$$\frac{dr}{dt} = \frac{dr}{d\theta}\frac{d\theta}{dt} = \left(-u^{-2}\frac{du}{d\theta}\right)\left(\ell u^2\right) = -\ell\frac{du}{d\theta}$$

and a second derivative of

$$\frac{d^2r}{dt^2} = \frac{d}{dt}\left(\frac{dr}{dt}\right) = \frac{d}{d\theta}\left(\frac{dr}{dt}\right)\frac{d\theta}{dt} = \frac{d}{d\theta}\left(-\ell\frac{du}{d\theta}\right)\left[\ell u^2\right] = -\ell^2u^2\frac{d^2u}{d\theta^2}.$$

If we make these substitutions in our equation of motion, we get

$$-\ell^2u^2\frac{d^2u}{d\theta^2} - \ell^2u^3 + GM_{\odot}u^2 = 0$$

and, dividing through by $-\ell^2u^2$, we get

$$\begin{aligned}\frac{d^2u}{d\theta^2} + u - \frac{GM_{\odot}}{\ell^2} &= 0\\ \frac{d^2u}{d\theta^2} + u &= \frac{GM_{\odot}}{\ell^2}.\end{aligned}$$

This has now become a very typical differential equation that we'll solve with another guess. Based on its form, the second derivative of $u(\theta)$ must be proportional to the same function as $u(\theta)$. This is only true for $\cos\theta$ and $\sin\theta$. Normally, we'd write the general solution as a linear combination of these specific ones, but it may be better in our case to just give the $\cos\theta$ a phase angle to accommodate the $\sin\theta$. Therefore, our general solution is in the form of

$$u(\theta) = \frac{GM_{\odot}}{\ell^2} + A\cos(\theta + \theta_0)\,.$$

Since the phase angle, θ_0, just determines orientation in this example, we can define it as zero giving us a general solution of

$$u(\theta) = \frac{GM_{\odot}}{\ell^2} + A\cos\theta$$

or, because $r = u^{-1}$,

$$r(\theta) = \frac{1}{\frac{GM_\odot}{\ell^2} + A\cos\theta}$$

$$r(\theta) = \frac{\frac{\ell^2}{GM_\odot}}{1 + B\cos\theta}.$$

This matches the form of the equation for conic sections. If we choose $r(0) = r_0$ (i.e. the perihelion) and $B = e$ (i.e. the eccentricity), then

$$\boxed{r(\theta) = \frac{r_0\,(1+e)}{1 + e\cos\theta}}, \tag{4.4.2}$$

which includes circles ($e = 0$), ellipses ($0 < e < 1$), parabolas ($e = 1$), and hyperbolas ($e > 1$). That makes our result a more generalized statement of Kepler's first law of planetary motion. This is exactly what we would expect if we're analyzing the motions of bodies in a gravitational field.

Example 4.4.3

A double pendulum is constructed as follows: A rigid string (of negligible mass) of length L_1 connects a mass m_1 to a perfectly rigid ceiling. Another rigid string (of negligible mass) of length L_2 connects another mass m_2 to the bottom of m_1.

1. The position of m_1 is represented by (r_1, θ_1) in cylindrical coordinates. Similarly, we can represent the position of m_2 by (r_2, θ_2) making the total set of generalized coordinates $(r_1, \theta_1, r_2, \theta_2)$. Four generalized coordinates could be a bit challenging, so let's see if we can simplify this with some constraints. We know from the example's wording the strings are rigid. This means they never bend or change length (i.e. $r_1 = L_1$). Even though $r_2 \neq L_2$, it can similarly be written in terms of just the lengths and the angles. Therefore, our best choice for the generalized coordinates, q_i, are (θ_1, θ_2).

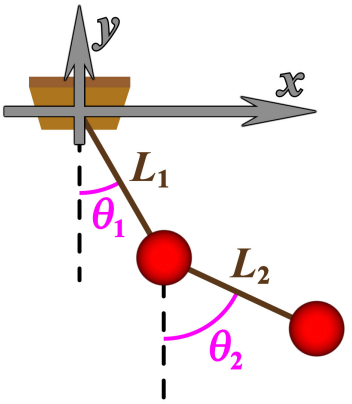


Figure 4.7: The strings are of constant length L_1 and L_2 and the pendulum bobs are free to swing in a two-dimension plane. The angle for each bob is measured from its respective vertical.

2. Based on Figure 4.7, we can write the coordinate transformations as

$$\left\{\begin{array}{lcl} x_1 &=& L_1 \sin\theta_1 \\ y_1 &=& -L_1 \cos\theta_1 \\ x_2 &=& x_1 + L_2 \sin\theta_2 = L_1 \sin\theta_1 + L_2 \sin\theta_2 \\ y_2 &=& y_1 - L_2 \cos\theta_2 = -L_1 \cos\theta_1 - L_2 \cos\theta_2 \end{array}\right\}$$

and the first time-derivatives are

$$\left\{\begin{array}{lcl} \dot{x}_1 &=& L_1\dot{\theta}_1 \cos\theta_1 \\ \dot{y}_1 &=& L_1\dot{\theta}_1 \sin\theta_1 \\ \dot{x}_2 &=& L_1\dot{\theta}_1 \cos\theta_1 + L_2\dot{\theta}_2 \cos\theta_2 \\ \dot{y}_2 &=& L_1\dot{\theta}_1 \sin\theta_1 + L_2\dot{\theta}_2 \sin\theta_2 \end{array}\right\}.$$

3. Our coordinate transformations make finding the potential and kinetic energy very straight forward. The potential energy is

$$\begin{array}{rcl} V &=& m_1 g h_1 + m_2 g h_2 = m_1 g y_1 + m_2 g y_2 \\ &=& -m_1 g L_1 \cos\theta_1 - m_2 g \left(L_1 \cos\theta_1 + L_2 \cos\theta_2\right) \\ &=& -\left(m_1 + m_2\right) g L_1 \cos\theta_1 - m_2 g L_2 \cos\theta_2 \end{array}$$

and kinetic energy is given by

$$K = \tfrac{1}{2} m_1 v_1^2 + \tfrac{1}{2} m_2 v_2^2 = \tfrac{1}{2} m_1 \left(\dot{x}_1^2 + \dot{y}_1^2\right) + \tfrac{1}{2} m_2 \left(\dot{x}_2^2 + \dot{y}_2^2\right).$$

We can see here that the kinetic energy of the system is going to become a very long equation, so it might be best to consider the two masses separately for now and bring them back together later. For m_1, we have

$$\begin{aligned} K_1 &= \frac{1}{2}m_1\left[\left(L_1\dot{\theta}_1\cos\theta_1\right)^2+\left(L_1\dot{\theta}_1\sin\theta_1\right)^2\right] \\ &= \frac{1}{2}m_1\left(L_1^2\dot{\theta}_1^2\cos^2\theta_1+L_1^2\dot{\theta}_1^2\sin^2\theta_1\right) \end{aligned}$$

and, since $\sin^2\theta+\cos^2\theta=1$,

$$K_1 = \tfrac{1}{2}m_1L_1^2\dot{\theta}_1^2$$

For m_2, we have

$$K_2 = \frac{1}{2}m_2\left[\left(L_1\dot{\theta}_1\cos\theta_1+L_2\dot{\theta}_2\cos\theta_2\right)^2+\left(L_1\dot{\theta}_1\sin\theta_1+L_2\dot{\theta}_2\sin\theta_2\right)^2\right]$$

$$\begin{aligned} K_2 &= \tfrac{1}{2}m_2\left(L_1^2\dot{\theta}_1^2\cos^2\theta_1+2L_1L_2\dot{\theta}_1\dot{\theta}_2\cos\theta_1\cos\theta_2+L_2^2\dot{\theta}_2^2\cos^2\theta_2\right. \\ &\quad \left.+L_1^2\dot{\theta}_1^2\sin^2\theta_1+2L_1L_2\dot{\theta}_1\dot{\theta}_2\sin\theta_1\sin\theta_2+L_2^2\dot{\theta}_2^2\sin^2\theta_2\right) \end{aligned}$$

and, since $\sin^2\theta+\cos^2\theta=1$ and $\cos A\cos B+\sin A\sin B=\cos(A-B)$,

$$K_2 = \tfrac{1}{2}m_2\left[L_1^2\dot{\theta}_1^2+2L_1L_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2)+L_2^2\dot{\theta}_2^2\right].$$

Bringing these back together to find the total kinetic energy, we get

$$\begin{aligned} K &= \tfrac{1}{2}m_1L_1^2\dot{\theta}_1^2+\tfrac{1}{2}m_2\left[L_1^2\dot{\theta}_1^2+2L_1L_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2)+L_2^2\dot{\theta}_2^2\right] \\ &= \tfrac{1}{2}\left(m_1+m_2\right)L_1^2\dot{\theta}_1^2+\tfrac{1}{2}m_2L_2^2\dot{\theta}_2^2+m_2L_1L_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2). \end{aligned}$$

4. The Lagrangian is

$$\mathcal{L} = K - V$$

$$\begin{aligned} \mathcal{L} &= \tfrac{1}{2}\left(m_1+m_2\right)L_1^2\dot{\theta}_1^2+\frac{1}{2}m_2L_2^2\dot{\theta}_2^2+m_2L_1L_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2) \\ &\quad -\left[-\left(m_1+m_2\right)gL_1\cos\theta_1-m_2gL_2\cos\theta_2\right] \end{aligned}$$

$$\begin{aligned} \mathcal{L} &= \tfrac{1}{2}\left(m_1+m_2\right)L_1^2\dot{\theta}_1^2+\tfrac{1}{2}m_2L_2^2\dot{\theta}_2^2+m_2L_1L_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2) \\ &\quad +\left(m_1+m_2\right)gL_1\cos\theta_1+m_2gL_2\cos\theta_2 \end{aligned}$$

5. Plugging this into Lagrange's equation, we get

$$\left\{\begin{array}{l} \dfrac{\partial \mathcal{L}}{\partial \theta_1} - \dfrac{d}{dt}\left(\dfrac{\partial \mathcal{L}}{\partial \dot{\theta}_1}\right) = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \theta_2} - \dfrac{d}{dt}\left(\dfrac{\partial \mathcal{L}}{\partial \dot{\theta}_2}\right) = 0 \end{array}\right\}$$

For clarity, we'll evaluate each term of each equation separately and then state it all together at the end. The terms of equation for θ_1 will be

$$\frac{\partial \mathcal{L}}{\partial \theta_1} = -m_2 L_1 L_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1 - \theta_2) - (m_1 + m_2)\, g L_1 \sin\theta_1$$

and

$$\begin{aligned} -\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{\theta}_1}\right) &= -\frac{d}{dt}\left[(m_1 + m_2)\, L_1^2 \dot{\theta}_1 + m_2 L_1 L_2 \dot{\theta}_2 \cos(\theta_1 - \theta_2)\right] \\ &= -(m_1 + m_2)\, L_1^2 \ddot{\theta}_1 - m_2 L_1 L_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) \\ &\quad + m_2 L_1 L_2 \dot{\theta}_2 \sin(\theta_1 - \theta_2)\left[\dot{\theta}_1 - \dot{\theta}_2\right] \\ &= -(m_1 + m_2)\, L_1^2 \ddot{\theta}_1 - m_2 L_1 L_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) \\ &\quad + m_2 L_1 L_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1 - \theta_2) - m_2 L_1 L_2 \dot{\theta}_2^2 \sin(\theta_1 - \theta_2)\,. \end{aligned}$$

If we cancel all like terms and divide through by $-L_1$, the common factor to all terms, we get

$$\begin{aligned} 0 &= (m_1 + m_2)\, L_1 \ddot{\theta}_1 + m_2 L_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) \\ &\quad + m_2 L_2 \dot{\theta}_2^2 \sin(\theta_1 - \theta_2) + (m_1 + m_2)\, g \sin\theta_1. \end{aligned}$$

Performing these same operations on the equation for θ_2, we get the terms

$$\frac{\partial \mathcal{L}}{\partial \theta_2} = m_2 L_1 L_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1 - \theta_2) - m_2 g L_2 \sin\theta_2$$

and

$$\begin{aligned} -\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{\theta}_2}\right) &= -\frac{d}{dt}\left[m_2 L_2^2 \dot{\theta}_2 + m_2 L_1 L_2 \dot{\theta}_1 \cos(\theta_1 - \theta_2)\right] \\ &= -m_2 L_2^2 \ddot{\theta}_2 - m_2 L_1 L_2 \ddot{\theta}_1 \cos(\theta_1 - \theta_2) \\ &\quad + m_2 L_1 L_2 \dot{\theta}_1 \sin(\theta_1 - \theta_2)\left[\dot{\theta}_1 - \dot{\theta}_2\right] \\ &= -m_2 L_2^2 \ddot{\theta}_2 - m_2 L_1 L_2 \ddot{\theta}_1 \cos(\theta_1 - \theta_2) \\ &\quad + m_2 L_1 L_2 \dot{\theta}_1^2 \sin(\theta_1 - \theta_2) - m_2 L_1 L_2 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_1 - \theta_2)\,. \end{aligned}$$

If we cancel all like terms and divide through by $-m_2 L_2$, the common factor to all terms, we get

$$0 = L_2 \ddot{\theta}_2 + L_1 \ddot{\theta}_1 \cos(\theta_1 - \theta_2) - L_1 \dot{\theta}_1^2 \sin(\theta_1 - \theta_2) + g \sin\theta_2.$$

Writing these together, we have a system of coupled second-order differential equations that represent the equations of motion for this system given by

$$\boxed{\begin{cases} 0 &= (m_1 + m_2)\, L_1 \ddot{\theta}_1 + m_2 L_2 \ddot{\theta}_2 \cos(\theta_1 - \theta_2) \\ &\quad + m_2 L_2 \dot{\theta}_2^2 \sin(\theta_1 - \theta_2) + (m_1 + m_2)\, g \sin\theta_1 \\ 0 &= L_2 \ddot{\theta}_2 + L_1 \ddot{\theta}_1 \cos(\theta_1 - \theta_2) - L_1 \dot{\theta}_1^2 \sin(\theta_1 - \theta_2) + g \sin\theta_2. \end{cases}}$$

These equations of motion cannot be solved analytically as was with Examples 4.4.1 and 4.4.2. However, a numerical method can be used to integrate them through a spreadsheet to attain a graphical solution (or even programmed into an animation). The most widely used (and my personal favorite) is the forth-order Runge-Kutta method given in Section A.1.

If we apply Runge-Kutta to the double pendulum, then we'll need to algebraically manipulate our equations of motion quite a bit ultimately arriving at

$$\begin{cases} \dot{\theta}_1 &= \omega_1 \\ \dot{\omega}_1 &= \dfrac{-\gamma\left[L_1\omega_1^2 \cos(\theta_1-\theta_2)+L_2\omega_2^2\right]\sin(\theta_1-\theta_2)+g[\gamma\sin(\theta_2)\cos(\theta_1-\theta_2)-\sin(\theta_1)]}{L_1[1-\gamma\cos^2(\theta_1-\theta_2)]} \\ \dot{\theta}_2 &= \omega_2 \\ \dot{\omega}_2 &= \dfrac{\left[\gamma L_2\omega_2^2 \cos(\theta_1-\theta_2)+L_1\omega_1^2\right]\sin(\theta_1-\theta_2)+g[\sin(\theta_1)\cos(\theta_1-\theta_2)-\sin(\theta_2)]}{L_2[1-\gamma\cos^2(\theta_1-\theta_2)]} \end{cases}$$

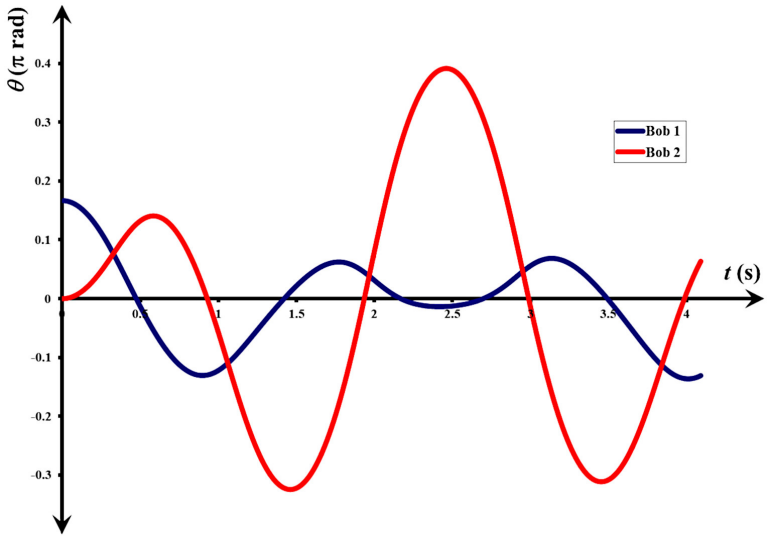


Figure 4.8: This graph shows both θ_1 and θ_2 as a function of time. The example is given for $\gamma = 0.2$, $L_1 = L_2 = 1$ m, $\theta_1(0) = \pi/6$, and $\theta_2(0) = 0$.

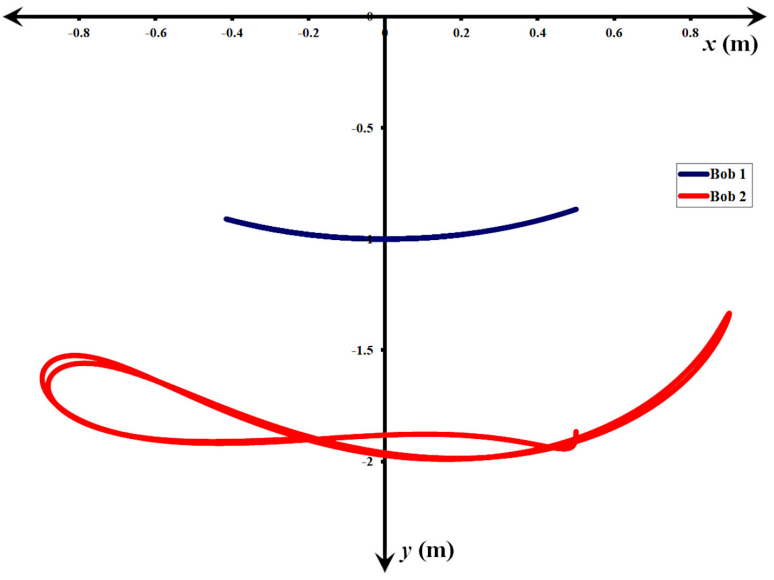


Figure 4.9: This is a representation of the path each pendulum bob has taken in space under the time interval given by Figure 4.8. The coordinate transformations have been used to convert back to x and y.

where $\gamma = m_2/(m_1 + m_2)$. The Runge-Kutta method should be applied to equations separately, but take note that $\dot{\omega}_1$ and $\dot{\omega}_2$ are dependent on all the following variables: ω_1, ω_2, θ_1, and θ_2. All of these variables also require initial conditions. The graphical solution is given by Figures 4.8 and 4.9.

4.5 Lagrange Multipliers

As seen in Section 4.4, Lagrange's equation is extremely useful when trying to find the equations of motion of a complex system. It can be done without concern for the forces involved and the process is roughly the same length regardless of the system. But what if we want to know something about the forces involved? Can Lagrange's equation help us then? The answer is most certainly yes, but it takes a bit of finesse.

The most efficient way to find equations of motion using Lagrange's equation is to incorporate the equations of constraint directly to reduce the number of generalized coordinates. In short, we made sure our generalized coordinates were completely independent so that they represented the degrees of freedom of the system. Unfortunately, when this is done, information is lost. The particular information being lost is the cause(s) of the constraint. We know from introductory physics that causes almost always involve forces. In this case, we call them **constraint forces.**

During the derivation in Section 4.2, we assumed the system was free of non-conservative forces which gave us Eq. 4.2.3. The only way to retain constraint forces is to relax our constraint (i.e. to not reduce our generalized coordinates). If we relax our constraint, then we must consider that the generalized coordinates are not all independent and Eq. 4.2.14 cannot be equal to zero. Which brings us to the next logical question: What is it equal to? The answer to this begins with considering the total force on our system is given by

$$\vec{F} = \vec{F}_{\text{conserv}} + \vec{F}_{\text{constraint}}. \tag{4.5.1}$$

The first terms can still be written in terms of potential energy just as before, but the second term needs some attention. We can use the method of **Lagrange Multipliers** to also write the constraint force as a gradient

resulting in a total force of

$$\vec{F} = -\vec{\nabla}V + \lambda\vec{\nabla}f \tag{4.5.2}$$

where $f(q_i) = 0$ is the equation of constraint written in terms of the generalized coordinates and λ is the Lagrange multiplier. It will become apparent later that the Lagrange multiplier is, in fact, equal to the constraint force.

Considering the system like this has added another unknown into the mix, but we also have one more Lagrange's equation than before since we haven't eliminated a generalized coordinate. The system is still solvable. If we carry the definition given in Eq. 4.5.2 through the derivation given in Section 4.2, then Eq. 4.2.4 becomes

$$\delta W = -\vec{\nabla}V \bullet \delta\vec{r} + \lambda\vec{\nabla}f \bullet \delta\vec{r}$$

and Eq. 4.2.5 becomes

$$\delta W = -\sum_{i=1}^{n}\frac{\partial V}{\partial q_i}\delta q_i + \lambda\sum_{i=1}^{n}\frac{\partial f}{\partial q_i}\delta q_i.$$

It turns out all we end up doing is carrying through a new term. Furthermore, the form of work given by Newton's second law is unaffected by the change. Therefore, we get

$$-\sum_{i=1}^{n}\frac{\partial V}{\partial q_i}\delta q_i + \lambda\sum_{i=1}^{n}\frac{\partial f}{\partial q_i}\delta q_i = \sum_{i=1}^{n}\left[\frac{d}{dt}\left(\frac{\partial K}{\partial \dot{q}_i}\right) - \frac{\partial K}{\partial q_i}\right]\delta q_i$$

$$\Rightarrow \sum_{i=1}^{n}\left[\frac{\partial\left(K-V\right)}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial K}{\partial \dot{q}_i}\right) + \lambda\frac{\partial f}{\partial q_i}\right]\delta q_i = 0.$$

$$\Rightarrow \sum_{i=1}^{n}\left[\frac{\partial\left(K-V\right)}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial\left(K-V\right)}{\partial \dot{q}_i}\right) + \lambda\frac{\partial f}{\partial q_i}\right]\delta q_i = 0.$$

Some texts prefer to move the λ term to the other side of the equation thereby answering our original question of what Lagrange's equation is equal to. However, I prefer to write as though it still is equal to zero and think of

λ as the constraint force that makes it zero. By the same processes as before, Eq. 4.2.14 now takes on the form

$$\frac{\partial \mathcal{L}}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right) + \lambda \frac{\partial f}{\partial q_i} = 0 \tag{4.5.3}$$

where $\mathcal{L} = K - V$ is the Lagrangian, λ is the constraint force (Lagrange multiplier), $f(q_i) = 0$ is the equation of constraint, q_i are the generalized coordinates, and $\dot{q}_i$ are the generalized velocities. Many situations involve more than one constraint force. If that is the case, then you can still solve by including a separate multiplier term (with a different multiplier) in Eq. 4.5.3 because each of these forces will involve their own equation of constraint. However, relaxing constraints can make a solution very long and tedious, so it may be better to solve the problem multiple times including each constraint one at a time.

4.6 Applications of Lagrange Multipliers

The process to solving these is very similar to that given in Section 4.4.

1. *Determine the set of generalized coordinates and the equation(s) of constraint for the system.* Remember to not reduce the coordinates fully. The appropriate coordinate will remain until we apply our equation of constraint at the very end.

2. *Write out the coordinate transformations.* In other words, write the cartesian coordinates of each object in terms of the generalized coordinates and take each of their first time-derivatives.

3. *Write out the potential and kinetic energy of the system in terms the generalized coordinates.* If you have multiple bodies in the system, then you can find the total by adding the corresponding energy from all the bodies together.

4. *Find the Lagrangian of the system.* Recall $\mathcal{L} = K - V$.

5. *Plug the Lagrangian into Lagrange's equation.* See Eq. 4.5.3.

Example 4.6.1

Returning to Example 4.4.1, find the constraint force causing the ball to roll without slipping.

1. As before, we will define x as the distance the ball has traveled down the incline and θ as the angle through which the ball has rotated. The equation of constraint for this example is $x = R\theta$ or $f(x, \theta) = x - R\theta = 0$. The y-direction may still be eliminated because the ball simply being constrained to the incline is caused by a different force.

2. Again, just as in Example 4.4.1, the coordinate transformations are unnecessary.

3. The potential and kinetic energy of the ball are given by

$$V = mgh = -mgx\sin\phi$$

and, since $I = \frac{2}{5}mR^2$ for a solid sphere, we get

$$\begin{aligned} K &= \tfrac{1}{2}mv^2 + \tfrac{1}{2}I\omega^2 = \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{2}I\dot{\theta}^2 \\ &= \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{5}mR^2\dot{\theta}^2 \end{aligned}$$

4. The Lagrangian is

$$\begin{aligned} \mathcal{L} &= K - V = \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{5}mR^2\dot{\theta}^2 - (-mgx\sin\phi) \\ &= \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{5}mR^2\dot{\theta}^2 + mgx\sin\phi. \end{aligned}$$

5. This time when we plug this into Lagrange's equation, there are two equations because there are two generalized coordinates. We get

$$\left\{\begin{aligned} \frac{\partial\mathcal{L}}{\partial x} - \frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{x}}\right) + \lambda\frac{\partial f}{\partial x} &= 0 \\ \frac{\partial\mathcal{L}}{\partial\theta} - \frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{\theta}}\right) + \lambda\frac{\partial f}{\partial\theta} &= 0 \end{aligned}\right\}$$

$$\left\{\begin{aligned} mg\sin\phi - \frac{d}{dt}\left(m\dot{x}\right) + \lambda &= 0 \\ 0 - \frac{d}{dt}\left(\frac{2}{5}mR^2\dot{\theta}\right) - \lambda R &= 0 \end{aligned}\right\}$$

We can take note here that if λ is a force acting on the outside edge of the ball, then λR is the torque that it causes. When you perform Lagrange's equation with respect to a distance, the terms that result are forces. When you perform Lagrange's equation with respect to an angle, the terms that result are torques. Simplifying a bit, we get

$$\left\{\begin{array}{rcl} mg\sin\phi - m\ddot{x} + \lambda &=& 0 \\ -\frac{2}{5}mR\ddot{\theta} - \lambda &=& 0 \end{array}\right\}$$

We are looking for λ, so lets start with the second equation and eliminate some of the other unknowns. Solving for $R\ddot{\theta}$, we get

$$\frac{2}{5}mR\ddot{\theta} = -\lambda$$
$$R\ddot{\theta} = -\frac{5\lambda}{2m}.$$

Since, from the equation of constraint, $x = R\theta \Rightarrow \ddot{x} = R\ddot{\theta}$, we get

$$\ddot{x} = -\frac{5\lambda}{2m}$$

and we can now eliminate $\ddot{x}$ from the first equation in our set resulting in

$$\begin{array}{rcl} mg\sin\phi - m\left(-\frac{5\lambda}{2m}\right) + \lambda &=& 0 \\ mg\sin\phi + \frac{5}{2}\lambda + \lambda &=& 0 \\ mg\sin\phi + \frac{7}{2}\lambda &=& 0 \end{array}$$

$$\boxed{\lambda = -\frac{2}{7}mg\sin\phi}.$$

This final answer is simply the force of static friction acting on the outside edge of the ball, which is exactly what we would expect and exactly the result we would find using Newton's laws.

Example 4.6.2

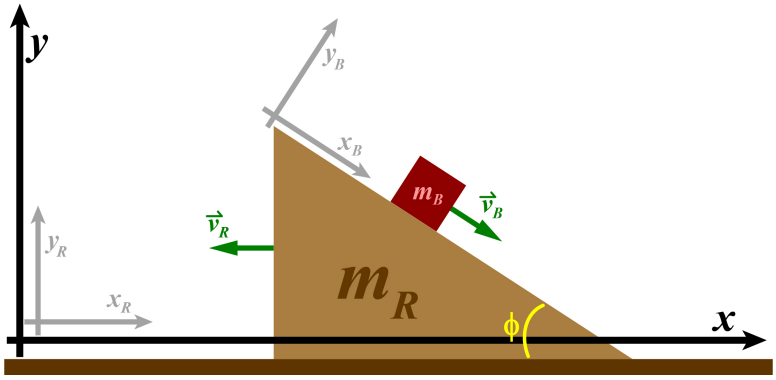


Figure 4.10: This figure shows the motions of the ramp and block as well as their respective coordinate systems. The coordinate transformations are a way to move (within the math) between these special coordinate systems and the universal xy system.

A block of mass m_B is sliding down a frictionless wedge-shaped ramp of mass m_R which also free to move along a frictionless horizontal surface. The sliding surface of the ramp makes an angle, ϕ, with the horizontal surface. Find the constraint force keeping the block on the wedge.

1. Based on Figure 4.10, we can see that the positions of the objects in the system are represented by (x_B, y_B, x_R, y_R). If we were just concerned with the equations of motion, then (x_B, x_R) would be enough since the wedge is constrained to the horizontal surface and the block is constrained to the ramp. However, if we want the force constraining the block to the ramp, then we need to keep y_B. Therefore, the generalize coordinates, q_i, are (x_B, y_B, x_R). The equation of constraint will be $f(x_B, y_B, x_R) = y_B = 0$.

2. We can write the coordinate transformations as

$$\left\{\begin{array}{rcl} x &=& x_R \\ y &=& 0 \\ x &=& x_R + x_B\cos\phi + y_B\sin\phi \\ y &=& -x_B\sin\phi + y_B\cos\phi \end{array}\right\}$$

where we have dropped the subscripts on the left side to emphasize that those coordinate are in the xy frame. Just keep in mind, the first two are for the ramp and the last two are for the block. The first

time-derivatives of the coordinate transformations can be written as

$$\left.\begin{array}{rcl} \dot{x} &=& \dot{x}_R \\ \dot{y} &=& 0 \\ \dot{x} &=& \dot{x}_R + \dot{x}_B\cos\phi + \dot{y}_B\sin\phi \\ \dot{y} &=& -\dot{x}_B\sin\phi + \dot{y}_B\cos\phi \end{array}\right\}.$$

It seems counter-intuitive that the block would have a velocity in the y_B direction. The easiest way to stay sane while conceptualizing all this is to remember you're carrying through a zero. However, at no point should you be plugging in a zero. This cannot be done until all the derivatives are taken and you have a set of equations of motion that include the constraint force.

3. Using our coordinate transformations, the potential energy is

$$\begin{array}{rcl} V &=& m_R g h_R + m_B g h_B = 0 + m_B g\left(-x_B\sin\phi + y_B\cos\phi\right) \\ &=& -m_B g\left(x_B\sin\phi - y_B\cos\phi\right) \end{array}$$

and kinetic energy is given by

$$K = \tfrac{1}{2}m_R v_R^2 + \tfrac{1}{2}m_B v_B^2 = \tfrac{1}{2}m_R\left(\dot{x}^2+\dot{y}^2\right)_R + \tfrac{1}{2}m_B\left(\dot{x}^2+\dot{y}^2\right)_B.$$

We can see here that the kinetic energy of the system is going to become an extremely long equation, so it might be best to consider the two objects separately for now and bring them back together later. For the ramp, we have

$$K_R = \tfrac{1}{2}m_R\left(\dot{x}_R^2 + 0^2\right) = \tfrac{1}{2}m_R\dot{x}_R^2$$

The kinetic energy of the block is where things get nasty. We get

$$K_B = \tfrac{1}{2}m_B\left[\left(\dot{x}_R + \dot{x}_B\cos\phi + \dot{y}_B\sin\phi\right)^2 + \left(-\dot{x}_B\sin\phi + \dot{y}_B\cos\phi\right)^2\right]$$

$$\begin{array}{rcl} K_B &=& \tfrac{1}{2}m_B\left(\dot{x}_R^2 + 2\dot{x}_R\dot{x}_B\cos\phi + 2\dot{x}_R\dot{y}_B\sin\phi\right. \\ && +\dot{x}_B^2\cos^2\phi + 2\dot{x}_B\dot{y}_B\sin\phi\cos\phi + \dot{y}_B^2\sin^2\phi \\ && \left.+\dot{x}_B^2\sin^2\phi - 2\dot{x}_B\dot{y}_B\sin\phi\cos\phi + \dot{y}_B^2\cos^2\phi\right) \end{array}$$

and, since $\sin^2\phi + \cos^2\phi = 1$,

$$K_B = \tfrac{1}{2}m_B\left(\dot{x}_R^2 + 2\dot{x}_R\dot{x}_B\cos\phi + 2\dot{x}_R\dot{y}_B\sin\phi + \dot{x}_B^2 + \dot{y}_B^2\right).$$

Bringing these back together to find the total kinetic energy, we get

$$\begin{aligned} K &= \tfrac{1}{2}m_R\dot{x}_R^2 + \tfrac{1}{2}m_B\left(\dot{x}_R^2 + 2\dot{x}_R\dot{x}_B\cos\phi + 2\dot{x}_R\dot{y}_B\sin\phi + \dot{x}_B^2 + \dot{y}_B^2\right) \\ &= \tfrac{1}{2}m_B\left(2\dot{x}_R\dot{x}_B\cos\phi + 2\dot{x}_R\dot{y}_B\sin\phi + \dot{x}_B^2 + \dot{y}_B^2\right) \\ &\quad + \tfrac{1}{2}\left(m_R + m_B\right)\dot{x}_R^2. \end{aligned}$$

4. The Lagrangian is

$$\begin{aligned} \mathcal{L} &= K - V \\ &= \tfrac{1}{2}m_B\left(2\dot{x}_R\dot{x}_B\cos\phi + 2\dot{x}_R\dot{y}_B\sin\phi + \dot{x}_B^2 + \dot{y}_B^2\right) \\ &\quad + \tfrac{1}{2}\left(m_R + m_B\right)\dot{x}_R^2 - \left[-m_Bg\left(x_B\sin\phi - y_B\cos\phi\right)\right] \\ &= \tfrac{1}{2}m_B\left(2\dot{x}_R\dot{x}_B\cos\phi + 2\dot{x}_R\dot{y}_B\sin\phi + \dot{x}_B^2 + \dot{y}_B^2\right) \\ &\quad + \tfrac{1}{2}\left(m_R + m_B\right)\dot{x}_R^2 + m_Bg\left(x_B\sin\phi - y_B\cos\phi\right). \end{aligned}$$

5. Plugging this into Lagrange's equation, we get

$$\left\{\begin{aligned} \frac{\partial\mathcal{L}}{\partial x_R} - \frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{x}_R}\right) + \lambda\frac{\partial f}{\partial x_R} &= 0 \\ \frac{\partial\mathcal{L}}{\partial x_B} - \frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{x}_B}\right) + \lambda\frac{\partial f}{\partial x_B} &= 0 \\ \frac{\partial\mathcal{L}}{\partial y_B} - \frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{y}_B}\right) + \lambda\frac{\partial f}{\partial y_B} &= 0 \end{aligned}\right\}$$

$$\left\{\begin{aligned} 0 - \frac{d}{dt}\left[\left(m_R + m_B\right)\dot{x}_R + m_B\left(\dot{x}_B\cos\phi + \dot{y}_B\sin\phi\right)\right] + 0 &= 0 \\ m_Bg\sin\phi - \frac{d}{dt}\left(m_B\dot{x}_R\cos\phi + m_B\dot{x}_B\right) + 0 &= 0 \\ -m_Bg\cos\phi - \frac{d}{dt}\left(m_B\dot{x}_R\sin\phi + m_B\dot{y}_B\right) + \lambda &= 0 \end{aligned}\right\}.$$

$$\left\{\begin{aligned} -\left(m_R + m_B\right)\ddot{x}_R - m_B\ddot{x}_B\cos\phi - m_B\ddot{y}_B\sin\phi &= 0 \\ m_Bg\sin\phi - m_B\ddot{x}_R\cos\phi - m_B\ddot{x}_B &= 0 \\ -m_Bg\cos\phi - m_B\ddot{x}_R\sin\phi - m_B\ddot{y}_B + \lambda &= 0 \end{aligned}\right\}.$$

If we divide through by $-(m_R + m_B)$ in the first equation and $-m_B$ in the second, then they become

$$\left\{\begin{array}{rcl} \ddot{x}_R + \gamma\ddot{x}_B\cos\phi + \gamma\ddot{y}_B\sin\phi &=& 0 \\ -g\sin\phi + \ddot{x}_R\cos\phi + \ddot{x}_B &=& 0 \\ -m_B g\cos\phi - m_B\ddot{x}_R\sin\phi - m_B\ddot{y}_B + \lambda &=& 0 \end{array}\right\}$$

where $\gamma = m_B/(m_R + m_B)$. Since all of the derivatives are taken, we can plug in our zero from the equation of constraint in all appropriate places and arrive at

$$\left\{\begin{array}{rcl} \ddot{x}_R + \gamma\ddot{x}_B\cos\phi &=& 0 \\ -g\sin\phi + \ddot{x}_R\cos\phi + \ddot{x}_B &=& 0 \\ -m_B g\cos\phi - m_B\ddot{x}_R\sin\phi + \lambda &=& 0 \end{array}\right\}.$$

All is well in the world again now that y_B is gone. Carrying y_B through the problem has resulted in an extra equation and the currently unknown constraint force, λ. With a little algebra, we should be able to find it. Logically, if we want λ, then we need to start with the third equation. That means

$$\lambda = m_B g\cos\phi + m_B\ddot{x}_R\sin\phi,$$

but we need $\ddot{x}_R$. The first equation contains this, but we'll need $\ddot{x}_B$. The second equation contains $\ddot{x}_B$ resulting in

$$\ddot{x}_B = g\sin\phi - \ddot{x}_R\cos\phi$$

Plugging this back into the first equation, we get

$$\begin{array}{rcl} \ddot{x}_R + \gamma\left(g\sin\phi - \ddot{x}_R\cos\phi\right)\cos\phi &=& 0 \\ \ddot{x}_R + \gamma g\sin\phi\cos\phi - \gamma\ddot{x}_R\cos^2\phi &=& 0 \\ \left(1 - \gamma\cos^2\phi\right)\ddot{x}_R + \gamma g\sin\phi\cos\phi &=& 0 \end{array}$$

$$\ddot{x}_R = \frac{-\gamma g\sin\phi\cos\phi}{1 - \gamma\cos^2\phi}.$$

Making our way back to λ, we find that

$$\lambda = m_B g\cos\phi + m_B\left(\frac{-\gamma g\sin\phi\cos\phi}{1 - \gamma\cos^2\phi}\right)\sin\phi$$

$$\boxed{\lambda = m_B g \left(\cos\phi - \frac{\gamma \sin^2\phi \cos\phi}{1 - \gamma \cos^2\phi} \right)}.$$

It may not be clear which force this is, so let's simplify a bit to get a feel for it. If we wanted to fix the ramp in place, then we would need to alter one of the quantities in λ. The easiest way to do this is to make the mass of the ramp very large so it remain nearly (inertially) unaffected by the block. This would result in $\gamma = 0$ and $\lambda = m_B g \cos\phi$. This force is just the normal force acting on the block due to the ramp. Therefore, our λ above is just the normal force due to a ramp that moves.

4.7 Non-Conservative Forces

The method of adding force terms at the beginning of the derivation presented Section 4.5 can also be used to generalize Lagrange equation for the inclusion of non-conservative forces (e.g. kinetic friction). Eq. 4.5.1 would be written as

$$\vec{F} = \vec{F}_{\text{conserv}} + \vec{F}_{\text{constraint}} + \vec{F}_{\text{non-conserv}}. \tag{4.7.1}$$

Starting from Eq. 4.7.1, we can see that Eq. 4.5.3 becomes

$$\frac{\partial \mathcal{L}}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right) + \lambda \frac{\partial f}{\partial q_i} + Q_i = 0 \tag{4.7.2}$$

where

$$Q_i = \sum_{j=1}^{3} F_j \frac{\partial r_j}{\partial q_i}$$

are the **generalized forces** that include all the non-conservative forces involved in the system transformed to the set of generalized coordinates.

Sometimes generalized forces can be written in terms of a velocity ($\dot{q}_i$) dependent potential energy. If this is the case, then they will become part of the Lagrangian and merge with the first two terms in Eq. 4.7.2. However, when dealing with non-conservative forces, it is usually best to concede to Newton's laws of motion for practical purposes.

Chapter 5

Electrodynamics

5.1 Introduction

The concepts of electricity and magnetism have been studied since Ancient Greece. In fact, there are records indicating Thales of Miletus was rubbing fur on amber around 600 BCE to generate an attractive force. The Ancient Greeks also had lodestone, a naturally occurring magnet made of a mineral now called magnetite. They came up with a wide variety of hypotheses, but very little progress was made in understanding why these phenomena occur. Scientific studies today are conducted using the scientific method, a rigorous process backed by experimental confirmation. In the middle-to-late 19^{th} century, it had become clear that classical mechanics (and, therefore, the Lagrangian mechanics of Chapter 4) was not sufficient to fully describe these phenomena and that another form of mechanics would be required to explain them.

5.2 Experimental Laws

The idea of charge in the middle 19^{th} century was defined using electric current. This, in turn, had been defined a century early by Benjamin Franklin as the flow of a positive fluid from his experiments with lightning. As we know today, the charge flowing in a conductor is negative electrons, not a fluid. However, the definition was sufficient at the time.

Figure 5.1: Charles Coulomb

Coulomb's Law

In 1784, Charles Coulomb was studying the effects of charged objects and their influence on one another. He published a relationship that governed the force exerted by one charged object on another. It had the form

$$\vec{F}_E = k_E \frac{q_1 q_2}{r^2} \hat{r}\,, \tag{5.2.1}$$

where q_1 and q_2 are the charges of the two objects, r is the distance between their centers, and k_E is a constant of proportionality with a value of 8.988×10^9 Nm^2/C^2. We call this **Coulomb's law**. This relationship is referred to as an inverse square law and, as you can see, bares a striking resemblance to Newton's universal law of gravitation,

$$\vec{F}_g = -G \frac{m_1 m_2}{r^2} \hat{r}, \tag{5.2.2}$$

published by Newton over a century before. The simple appearance of Eqs. 5.2.1 and 5.2.2 is very useful when trying to understand the relationships between quantities. It is sometimes more useful in practical situations to write Eq. 5.2.1 in terms of position vectors,

$$\left\{ \begin{aligned} \vec{F}_{E12} &= k_E \frac{q_1 q_2}{|\vec{r}_1 - \vec{r}_2|^3} (\vec{r}_1 - \vec{r}_2) \\ \vec{F}_{E21} &= k_E \frac{q_1 q_2}{|\vec{r}_2 - \vec{r}_1|^3} (\vec{r}_2 - \vec{r}_1) \end{aligned} \right\}, \tag{5.2.3}$$

where $\vec{r}_1$ and $\vec{r}_2$ are the positions of q_1 and q_2, respectively. We have used

$$\hat{r} = \frac{\vec{r}}{r} \tag{5.2.4}$$

to eliminate the unit vector $\hat{r}$. The subscript of 12 indicates this is the force on q_1 due to q_2 and 21 the reverse. However, these equations lack the elegance found in Eq. 5.2.1.

Another limitation of both Eqs. 5.2.1 and 5.2.3 is that they only apply when the objects in question can be approximated as nearly stationary point charges. Furthermore, situations can arise where we may not know much about some of the charge involved due to system complexity. It is astronomically more useful to define a quantity known as a **field**. In this case, we'd call it an electric field (abbreviated as E-field). This field is a representation of how electric charge affects the surrounding space. Essentially, we're creating a mathematical middle-man. I realize, at first glance, it might seem more complicated to consider an entirely new quantity, but this E-field has incredible power (pardon the pun). We can determine the E-field around a charged object, whatever the shape, and then forget about that object when predicting its effect on a new charge in the region, as long as this new charge is small compared to the original so as to not affect its E-field. We can also measure the E-field in a region while never considering its source.

Starting with Eq. 5.2.1, we can write the basic definition of an E-field as

$$\vec{E} = k_E \frac{q}{r^2}\hat{r}\,, \tag{5.2.5}$$

where q is the charge generating the E-field. The electric force on a new charge, q_0, is then just $\vec{F}_E = q_0\vec{E}$. Based on this, we can also conceptualize an E-field as a measure of one charge's ability to exert a force on another charge. Again, however, Eq. 5.2.5 still only applies to charges which are approximately points.

To find an E-field due to a charge distribution, we can write Eq. 5.2.5 as

$$d\vec{E} = k_E \frac{dq}{r^2}\hat{r}, \tag{5.2.6}$$

where dq represents an infinitesimal portion of the charge distribution (i.e. charge element) dependent on $\vec{r}$ (i.e. both r and $\hat{r}$) and $d\vec{E}$ is the E-field element generated by dq. The value of r now represents the distance from dq to the point in space that is of interest. Nothing need be at that point, however, because we're only discussing how the charge distribution affects space itself. The total field can be found through superposition by integration (which is just a sum of an infinite number of infinitesimally small terms).

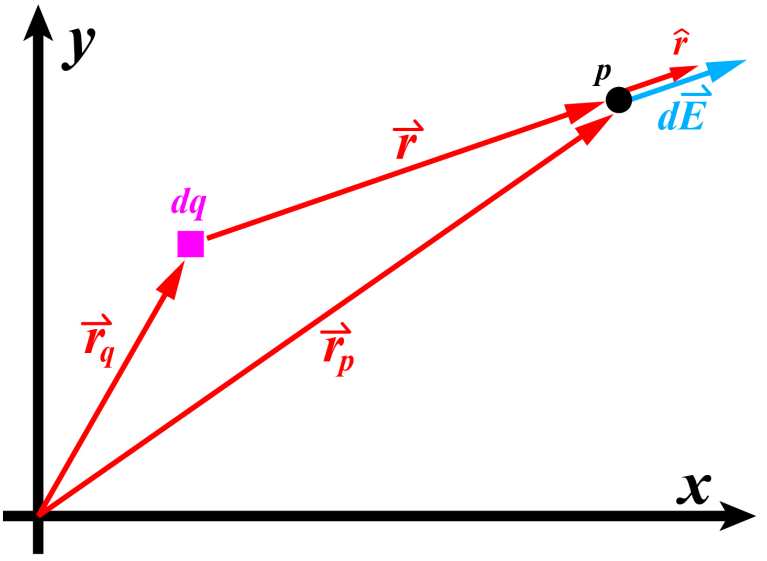


Figure 5.2: This diagram shows all the quantities used in Eqs. 5.2.6 and 5.2.7 in an arbitrary coordinate system. We can see clearly here that $\vec{r} = \vec{r}_p - \vec{r}_q$ because $\vec{r}_p = \vec{r}_q + \vec{r}$.

Writing Eq. 5.2.6 in terms of position vectors results in

$$d\vec{E} = k_E \frac{dq}{|\vec{r}_p - \vec{r}_q|^3} (\vec{r}_p - \vec{r}_q) \tag{5.2.7}$$

where $\vec{r}_p$ is the position of the point in space and $\vec{r}_q$ is the position of dq. We have used Eq. 5.2.4 to eliminate the unit vector. Once again, we lose elegance, but gain practical usefulness.

Just as with problems in Chapter 4, there is a methodical process for solving problems like this.

1. *Chose an arbitrary dq and find its value in terms of some spatial variable(s).* If you've positioned your coordinate system wisely, this should look relatively simple.

2. *Find $\vec{r}_p$ and $\vec{r}_q$ for the system.* This shouldn't be too difficult if you've drawn a good picture with the proper labels similar to Figure 5.2.

3. *Find $\vec{r} = \vec{r}_p - \vec{r}_q$ and $r = |\vec{r}_p - \vec{r}_q|$.* This takes the guesswork out of finding $\vec{r}$.

4. *Substitute from the previous step into Eq. 5.2.7 and separate into vector component terms.* In order to do the next step, these vector components should have constant directions. The Cartesian coordinate system is a common choice.

5. *Integrate over whatever variable(s) dq is dependent on.* Depending on the charge distribution this could be 1, 2, or 3 spatial variables.

Example 5.2.1

Find the electric field at an arbitrary point p in the space around a uniformly charged amber rod.

1. Based on the coordinate system chosen in Figure 5.3, we have charge distributed uniformly along the x-axis. Therefore,

$$\lambda = \frac{dq}{dx_q} = \text{constant} \quad \Rightarrow dq = \lambda \, dx_q$$

   where λ is the linear charge density. It is constant because the distribution is uniform. Uniformity is not a requirement in general, but a different distribution would certainly make the rest of this example rather complicated.

2. The point p chosen is arbitrary, but will remain constant through the following derivation because we're integrating along dq (i.e. the rod). The position of dq is also arbitrary so that we don't make any premature judgements about the form of $\vec{r}_q$. Figure 5.3 shows the two position vectors to clearly be

$$\vec{r}_p = x_p\hat{x} + y_p\hat{y}$$

   and

$$\vec{r}_q = x_q\hat{x}$$

   where x_q represents the variable of integration and we have suppressed the z-component through cylindrical symmetry about the x-axis (don't worry, we'll put it back in later). We should note the only circumstance in which $\vec{r}_q$ is constant is when there is no charge distribution at all, but simply a point charge.

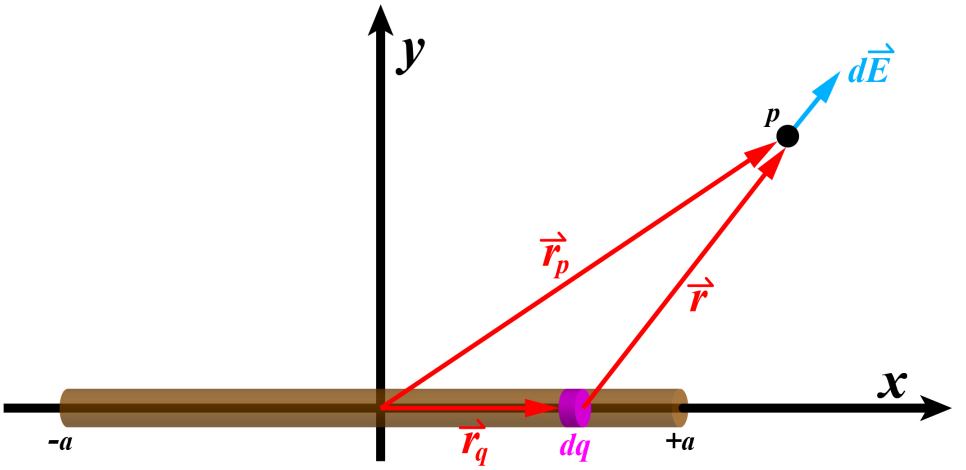


Figure 5.3: The amber rod is placed along the x-axis and all vectors from Eq. 5.2.7 are shown.

3. The vector $\vec{r}$ would be

$$\vec{r}_p - \vec{r}_q = (x_p\hat{x} + y_p\hat{y}) - (x_q\hat{x}) = (x_p - x_q)\,\hat{x} + y_p\hat{y}$$

which means r is

$$|\vec{r}_p - \vec{r}_q| = \sqrt{(x_p - x_q)^2 + y_p^2} = \left[(x_p - x_q)^2 + y_p^2\right]^{\frac{1}{2}}.$$

4. If we substitute these into Eq. 5.2.7, then we have

$$d\vec{E} = k_E \frac{\lambda dx_q}{\left[(x_p - x_q)^2 + y_p^2\right]^{3/2}} \left[(x_p - x_q)\,\hat{x} + y_p\hat{y}\right]$$

$$d\vec{E} = k_E\lambda\hat{x}\frac{(x_p - x_q)\,dx_q}{\left[(x_p - x_q)^2 + y_p^2\right]^{3/2}} + k_E\lambda\hat{y}\frac{y_p dx_q}{\left[(x_p - x_q)^2 + y_p^2\right]^{3/2}}.$$

5. If we want the total E-field due to the amber rod, we must integrate over all possible dq's. Using $\vec{E} = \int_0^{\vec{E}} d\vec{E}$, we get

$$\vec{E} = k_E\lambda\hat{x}\int_{-a}^{a}\frac{(x_p - x_q)\,dx_q}{\left[(x_p - x_q)^2 + y_p^2\right]^{3/2}} + k_E\lambda\hat{y}\int_{-a}^{a}\frac{y_p dx_q}{\left[(x_p - x_q)^2 + y_p^2\right]^{3/2}}.$$

From this point on, it will be a bit more clear if we discuss the components separately. If we define $\vec{E}$ as $E_x\hat{x} + E_y\hat{y}$, then

$$\left\{\begin{array}{rcl} E_x & = & k_E\lambda \displaystyle\int_{-a}^{a} \frac{(x_p - x_q)\, dx_q}{\left[(x_p - x_q)^2 + y_p^2\right]^{3/2}} \\ \\ E_y & = & k_E\lambda \displaystyle\int_{-a}^{a} \frac{y_p dx_q}{\left[(x_p - x_q)^2 + y_p^2\right]^{3/2}} \end{array}\right\}.$$

We can evaluate the x-component integral using a change of variable (something the mathematicians like to call a u-substitution). Choosing how to define the new variable is a bit of an art, but the desired result is always the same: make the integrand as simple as possible. This is done by choosing a definition for the new variable that is as complex as possible such that all forms of the old variable can vanish. In this case,

$$u = (x_p - x_q)^2 + y_p^2$$

would be the best choice. The first derivative of this is

$$\frac{du}{dx_q} = 2\,(x_p - x_q)\,(-1) = -2\,(x_p - x_q)$$

$$\Rightarrow (x_p - x_q)\, dx_q = -\frac{1}{2}du.$$

This results in an x-component of

$$E_x = k_E\lambda \int_{u_1}^{u_2} \frac{-1/2}{u^{3/2}}du = -\frac{1}{2}k_E\lambda \int_{u_1}^{u_2} u^{-3/2}du$$

$$E_x = -\frac{1}{2}k_E\lambda \left(\frac{u^{-1/2}}{-1/2}\right)\Bigg|_{u_1}^{u_2} = k_E\lambda \left(\frac{1}{u^{1/2}}\right)\Bigg|_{u_1}^{u_2}.$$

We can now transform back into the old variable x arriving at

$$E_x = k_E\lambda \left(\frac{1}{\sqrt{(x_p - x_q)^2 + y_p^2}}\right)\Bigg|_{-a}^{a}$$

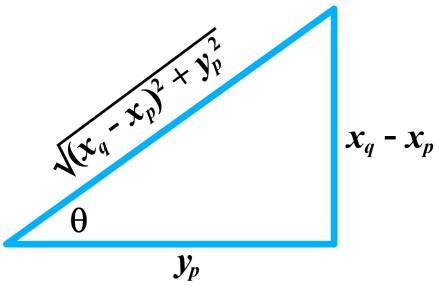


Figure 5.4: This reference triangle is used to transform the integrand of the y-component in Example 5.2.1. The side opposite the angle θ is labeled $x_q - x_p$ rather than $x_p - x_q$ to eliminate a negative sign from the transformation. This is mathematically legal because $(x_q - x_p)^2 = (x_p - x_q)^2$.

$$E_x = k_E \lambda \left(\frac{1}{\sqrt{(x_p - a)^2 + y_p^2}} - \frac{1}{\sqrt{(x_p + a)^2 + y_p^2}} \right).$$

Integrating the y-component is a bit trickier because the derivative of any u we chose isn't found in the integrand. We need to take advantage of a much more powerful change of variable: a trigonometric substitution (or trig-substitution). This involves using a reference triangle to change to an angular variable. Trig-substitutions only work on integrals involving square roots of squared terms analogous to Pythagorean theorem. Based on Figure 5.4, we have

$$\frac{x_q - x_p}{y_p} = \tan\theta \quad \Rightarrow x_q = y_p \tan\theta + x_p$$

and a first derivative of

$$\frac{dx_q}{d\theta} = y_p \sec^2\theta \quad \Rightarrow dx_q = y_p \sec^2\theta \, d\theta.$$

Rather than substituting our form for x_q into the integrand like the mathematicians would do, we can manipulate the integrand a bit to save us some time. We can perform something I like to call *voodoo math* (with a little foresight; we can add zeros, multiply by ones, add and subtract constants, etc. to simplify a mathematical expression). By multiplying the integrand by y_p^2 (a constant) and dividing by the same value outside the integral, we

arrive at

$$E_y = \frac{k_E \lambda}{y_p^2} \int_{-a}^{a} \frac{y_p^3}{\left[(x_p - x_q)^2 + y_p^2 \right]^{3/2}} dx_q.$$

Suddenly, with another look at Figure 5.4, our integrand simply becomes $\cos^3 \theta$ and E_y becomes

$$E_y = \frac{k_E \lambda}{y_p^2} \int_{\theta_1}^{\theta_2} \cos^3 \theta \left(y_p \sec^2 \theta \, d\theta \right)$$

$$E_y = \frac{k_E \lambda}{y_p} \int_{\theta_1}^{\theta_2} \cos \theta \, d\theta = \frac{k_E \lambda}{y_p} \left(\sin \theta \right) \Big|_{\theta_1}^{\theta_2}.$$

We can now transform back into the old variable x_q arriving at

$$E_y = \frac{k_E \lambda}{y_p} \left(\frac{-(x_p - x_q)}{\sqrt{(x_p - x_q)^2 + y_p^2}} \right) \Bigg|_{-a}^{a}$$

$$E_y = \frac{k_E \lambda}{y_p} \left(\frac{-(x_p - a)}{\sqrt{(x_p - a)^2 + y_p^2}} - \frac{-(x_p + a)}{\sqrt{(x_p + a)^2 + y_p^2}} \right)$$

$$E_y = \frac{k_E \lambda}{y_p} \left(\frac{x_p + a}{\sqrt{(x_p + a)^2 + y_p^2}} - \frac{x_p - a}{\sqrt{(x_p - a)^2 + y_p^2}} \right)$$

In summary, we can write the electric field as $\vec{E} = E_x \hat{x} + E_y \hat{y}$ where the components are

$$\left\{ \begin{array}{rcl} E_x & = & k_E \lambda \left(\dfrac{1}{\sqrt{(x_p - a)^2 + y_p^2}} - \dfrac{1}{\sqrt{(x_p + a)^2 + y_p^2}} \right) \\ E_y & = & \dfrac{k_E \lambda}{y_p} \left(\dfrac{x_p + a}{\sqrt{(x_p + a)^2 + y_p^2}} - \dfrac{x_p - a}{\sqrt{(x_p - a)^2 + y_p^2}} \right) \end{array} \right\}.$$

This result, because the point p was completely arbitrary, applies to all space around the rod. With that in mind, we can simplify things by dropping the p subscript for future uses. The components become

$$\left\{\begin{array}{rcl} E_x &=& k_E\lambda\left(\dfrac{1}{\sqrt{(x-a)^2+y^2}}-\dfrac{1}{\sqrt{(x+a)^2+y^2}}\right) \\ E_y &=& \dfrac{k_E\lambda}{y}\left(\dfrac{x+a}{\sqrt{(x+a)^2+y^2}}-\dfrac{x-a}{\sqrt{(x-a)^2+y^2}}\right) \end{array}\right\}.$$

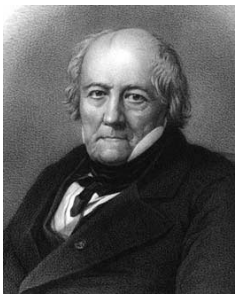
Furthermore, as mentioned in step 2, this system has cylindrical symmetry about the x-axis. This means the y-component could just as easily measure the distance from the rod in any direction perpendicular to the length of the rod. We can then transform the Cartesian y and z coordinates into a form of cylindrical coordinates (described in Section 1.2), s and ϕ, such that

$$\vec{s} = y\hat{y} + z\hat{z} = s\cos\phi\,\hat{y} + s\sin\phi\,\hat{z} = s\hat{s}.$$

This is slightly different from the standard definition only because of the orientation of the rod along the x-axis. In the xy-plane, the z-direction is equivalent to the ϕ-direction. We now have

$$\boxed{\left\{\begin{array}{rcl} E_x &=& k_E\lambda\left(\dfrac{1}{\sqrt{(x-a)^2+s^2}}-\dfrac{1}{\sqrt{(x+a)^2+s^2}}\right) \\ E_s &=& \dfrac{k_E\lambda}{s}\left(\dfrac{x+a}{\sqrt{(x+a)^2+s^2}}-\dfrac{x-a}{\sqrt{(x-a)^2+s^2}}\right) \\ E_\phi &=& 0 \end{array}\right\}} \tag{5.2.8}$$

where $\vec{E} = E_x\hat{x} + E_s\hat{s} + E_\phi\hat{\phi}$. This represents the completely general solution under the generalized coordinates (x, s, ϕ).


Jean-Baptiste Biot


Pierre-Simon Laplace

Figure 5.5: These people were important in the development the Biot-Savart law.

Biot-Savart Law

A somewhat similar relationship to Eq. 5.2.6 was discovered for magnetic fields, but it wouldn't arrive for almost another 40 years. Together in 1820, Jean-Baptiste Biot and Félix Savart announced they had discovered the magnetic force due to a current carrying conductor was proportional to $1/R$ and this force was perpendicular to the wire. This wasn't much of a result, but it was a start.

A mathematician named Pierre-Simon Laplace very quickly generalized this result in terms of a magnetic field $\vec{B}$, much like the electric field. Laplace's equation looked something like

$$d\vec{B} = k_M \frac{I d\vec{l} \times \hat{r}}{r^2}, \tag{5.2.9}$$

where I is a steady electric current generating the B-field, $d\vec{l}$ is the infinitesimal section of the conductor in the direction of the current, r is the distance between $Id\vec{l}$ and the point in space being examined, $\hat{r}$ is the unit vector in the direction of $\vec{r}$, and k_M is a constant of proportionality with a value of 1.0×10^{-7} N/A^2. This is what we now call the **Biot-Savart law**. The cross product in Eq. 5.2.9 indicates that $d\vec{B}$ is perpendicular to both $Id\vec{l}$ and $\hat{r}$ making it consistent with Biot and Savart's result. The vector sign is usually placed on the dl rather than I to emphasize the current is a steady, but it can really be placed on either.

We can generalize Eq. 5.2.9 much like we did with Eq. 5.2.7 resulting in

$$d\vec{B} = k_M \frac{I d\vec{l} \times (\vec{r}_p - \vec{r}_I)}{|\vec{r}_p - \vec{r}_I|^3}, \tag{5.2.10}$$

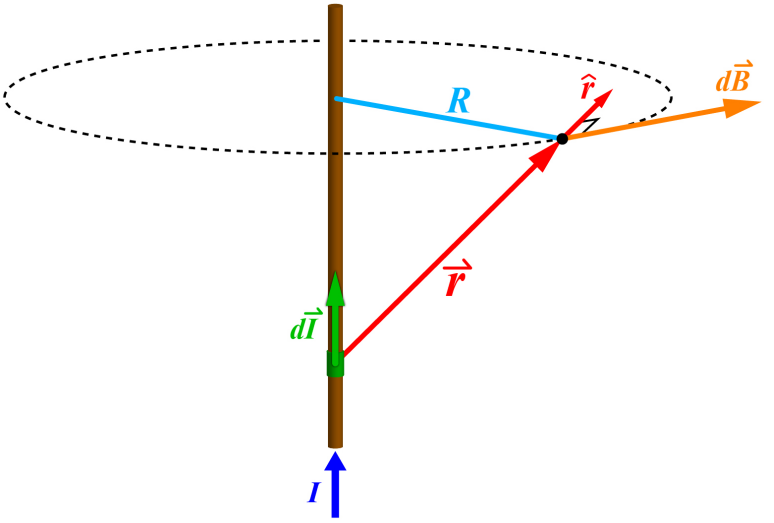


Figure 5.6: This diagram shows all the quantities used in Eq. 5.2.9 as well as the quantity R defined by Biot and Savart's discovery. The quantity $d\vec{B}$ is indicated as perpendicular to $\hat{r}$. It is also tangent to the dashed circle indicating it is also perpendicular to $d\vec{l}$.

where $\vec{r}_p$ is the position of the point in space and $\vec{r}_I$ is the position of $Id\vec{l}$. Again, we have used Eq. 5.2.4 to eliminate the unit vector just like we did with Coulomb's law. The methodical process for solving problems with the Biot-Savart law is similar to that of Coulomb's law.

1. *Chose an arbitrary $Id\vec{l}$ and find its value in terms of some spatial variable(s).* If you've positioned your coordinate system wisely, this should look relatively simple.

2. *Find $\vec{r}_p$ and $\vec{r}_I$ for the system.* This shouldn't be too difficult if you've drawn a good picture with the proper labels similar to Figure 5.6.

3. *Find $\vec{r} = \vec{r}_p - \vec{r}_I$ and $r = |\vec{r}_p - \vec{r}_I|$.* This takes the guesswork out of finding $\vec{r}$.

4. *Perform the cross product given by $Id\vec{l} \times (\vec{r}_p - \vec{r}_I)$.* This will save you writing time and keep things clear in your solution.

5. *Substitute from the previous step into Eq. 5.2.10 and separate into vector component terms.* In order to do the next step, these vector com-

ponents should have constant directions. The Cartesian coordinate system is a common choice.

6. *Integrate over whatever variable $I d\vec{l}$ is dependent on.* The form given in Eq. 5.2.10 is over a single variable, but it can be generalized to more. The quantity $I d\vec{l}$ is simply replaced by $\vec{K} dA_I$ (two variables) or $\vec{J} dV_I$ (three variables) depending on the type of electric current distribution.

Example 5.2.2

A circular conductor with a radius of R is carrying a steady current I. Find the magnetic field at an arbitrary point p around this loop.

1. Based on the coordinate system chosen in Fiqure 5.7, we have an electric current distribution in the xy-plane in the $\hat{\phi}$ direction. Therefore, we can write

$$I d\vec{l} = IR\, d\vec{\phi}_I = IR\, d\phi_I\, \hat{\phi}_I = IR\, d\phi_I\, (-\sin\phi_I\, \hat{x} + \cos\phi_I\, \hat{y})$$

   where we have taken advantage of Eq. 1.2.3 to write this in terms of vectors with constant direction.

2. The point p chosen is arbitrary, but will remain constant through the following derivation because we're integrating along $I d\vec{l}$ (i.e. the loop). For mathematical simplicity, however, we can suppress the y-component through cylindrical symmetry about the z-axis (don't worry, we'll put it back in later). Figure 5.7 shows

$$\vec{r}_p = x_p\hat{x} + y_p\hat{y} + z_p\hat{z} = x_p\hat{x} + z_p\hat{z}.$$

   The position of $I d\vec{l}$ is also arbitrary so that we don't make any premature judgements about the form of $\vec{r}_I$. Figure 5.7 shows

$$\vec{r}_I = R\hat{s}_I = R\,(\cos\phi_I\, \hat{x} + \sin\phi_I\, \hat{y})$$

   where we have taken advantage of Eq. 1.2.3 to write this in terms of vectors with constant direction.

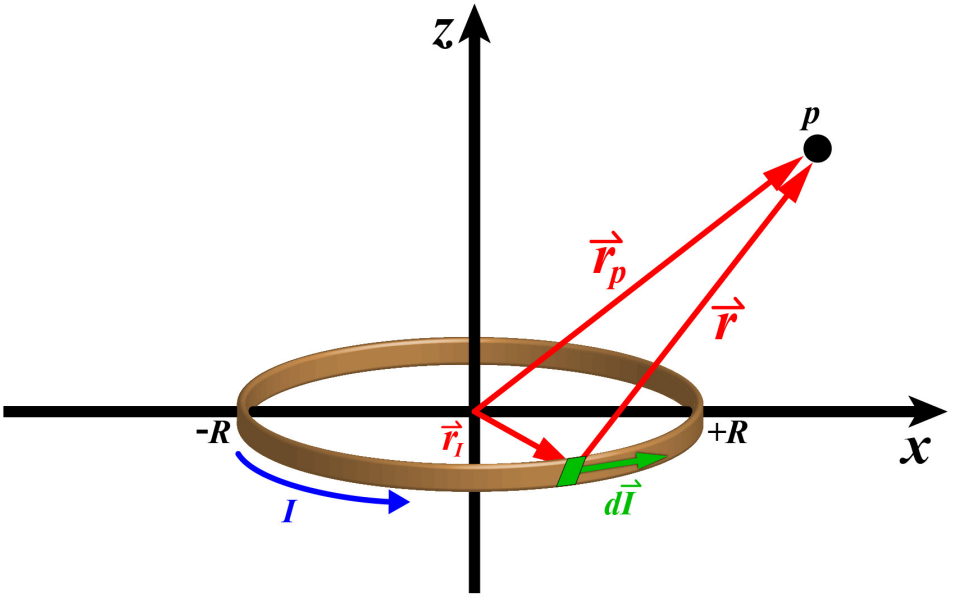


Figure 5.7: The conducting loop is placed in the xy-plane centered at the origin and all vectors from Eq. 5.2.10 are shown.

3. The vector $\vec{r}$ would be

$$\vec{r}_p - \vec{r}_I = (x_p\hat{x} + z_p\hat{z}) - R\left(\cos\phi_I\,\hat{x} + \sin\phi_I\,\hat{y}\right)$$

$$\vec{r}_p - \vec{r}_I = \left(x_p - R\cos\phi_I\right)\hat{x} + \left(-R\sin\phi_I\right)\hat{y} + z_p\hat{z}.$$

The integration we'll be doing later on will be easier if the quantities involved are unitless, so we'll define these: $x_R \equiv x_p/R$ and $z_R \equiv z_p/R$. Now, $\vec{r}$ can be written as

$$\vec{r}_p - \vec{r}_I = R\left[\left(x_R - \cos\phi_I\right)\hat{x} + \left(-\sin\phi_I\right)\hat{y} + z_R\hat{z}\right].$$

This means r is

$$|\vec{r}_p - \vec{r}_I| = R\sqrt{\left(x_R - \cos\phi_I\right)^2 + \left(-\sin\phi_I\right)^2 + z_R^2}$$

$$|\vec{r}_p - \vec{r}_I| = R\sqrt{x_R^2 - 2x_R\cos\phi_I + \cos^2\phi_I + \sin^2\phi_I + z_R^2}.$$

With a little rearranging and the trig identity $\cos^2\phi + \sin^2\phi = 1$, we get

$$|\vec{r}_p - \vec{r}_I| = R\sqrt{x_R^2 + z_R^2 + 1 - 2x_R\cos\phi_I}$$

$$|\vec{r}_p - \vec{r}_I| = R\left(x_R^2 + z_R^2 + 1 - 2x_R\cos\phi_I\right)^{1/2}$$

4. Using Eq. 2.2.4, the cross product in the integrand is

$$I\vec{dl} \times \vec{r} = I\vec{dl} \times (\vec{r}_p - \vec{r}_I)$$

$$= IR\, d\phi_I\,(-\sin\phi_I\,\hat{x} + \cos\phi_I\,\hat{y}) \times R\left[(x_R - \cos\phi_I)\,\hat{x} - \sin\phi_I\,\hat{y} + z_R\hat{z}\right]$$

$$= IR^2\, d\phi_I\ \det\begin{bmatrix} \hat{x} & \hat{y} & \hat{z} \\ -\sin\phi_I & \cos\phi_I & 0 \\ (x_R - \cos\phi_I) & -\sin\phi_I & z_R \end{bmatrix}$$

$$= IR^2\, d\phi_I\left(z_R\cos\phi_I\hat{x} + z_R\sin\phi_I\hat{y} + \left[\sin^2\phi_I - \cos\phi_I\,(x_R - \cos\phi_I)\right]\hat{z}\right)$$

$$= IR^2\, d\phi_I\left[z_R\cos\phi_I\hat{x} + z_R\sin\phi_I\hat{y} + \left(\sin^2\phi_I - x_R\cos\phi_I + \cos^2\phi_I\right)\hat{z}\right].$$

Again $\cos^2\phi + \sin^2\phi = 1$, so

$$I\vec{dl} \times \vec{r} = IR^2\, d\phi_I\left[z_R\cos\phi_I\hat{x} + z_R\sin\phi_I\hat{y} + (1 - x_R\cos\phi_I)\,\hat{z}\right].$$

5. If we substitute these into Eq. 5.2.10, then we have

$$d\vec{B} = k_M\frac{IR^2\, d\phi_I\left[z_R\cos\phi_I\hat{x} + z_R\sin\phi_I\hat{y} + (1 - x_R\cos\phi_I)\,\hat{z}\right]}{R^3\left(x_R^2 + z_R^2 + 1 - 2x_R\cos\phi_I\right)^{3/2}}$$

$$d\vec{B} = \frac{k_M I}{R}\left[\frac{z_R\cos\phi_I\hat{x} + z_R\sin\phi_I\hat{y} + (1 - x_R\cos\phi_I)\,\hat{z}}{\left(x_R^2 + z_R^2 + 1 - 2x_R\cos\phi_I\right)^{3/2}}\right]d\phi_I.$$

From this point on, it will be a bit more clear if we discuss the components separately. If we define $d\vec{B}$ and $dB_x\hat{x} + dB_y\hat{y} + dB_z\hat{z}$, then

$$\left\{\begin{array}{rcl} dB_x &=& \dfrac{k_M I}{R}\left[\dfrac{z_R\cos\phi_I\, d\phi_I}{\left(x_R^2+z_R^2+1-2x_R\cos\phi_I\right)^{3/2}}\right] \\ dB_y &=& \dfrac{k_M I}{R}\left[\dfrac{z_R\sin\phi_I\, d\phi_I}{\left(x_R^2+z_R^2+1-2x_R\cos\phi_I\right)^{3/2}}\right] \\ dB_z &=& \dfrac{k_M I}{R}\left[\dfrac{(1-x_R\cos\phi_I)\; d\phi_I}{\left(x_R^2+z_R^2+1-2x_R\cos\phi_I\right)^{3/2}}\right] \end{array}\right\}.$$

6. If we want the total B-field due to the conducting loop, we must integrate over all possible $d\phi_I$'s. Using $\vec{B} = \int_0^{\vec{B}} d\vec{B}$, we get

$$\left\{\begin{array}{rcl} B_x &=& \dfrac{k_M I}{R}\displaystyle\int_0^{2\pi}\dfrac{z_R\cos\phi\, d\phi_I}{\left(x_R^2+z_R^2+1-2x_R\cos\phi_I\right)^{3/2}} \\ B_y &=& \dfrac{k_M I}{R}\displaystyle\int_0^{2\pi}\dfrac{z_R\sin\phi\, d\phi_I}{\left(x_R^2+z_R^2+1-2x_R\cos\phi_I\right)^{3/2}} \\ B_z &=& \dfrac{k_M I}{R}\displaystyle\int_0^{2\pi}\dfrac{(1-x_R\cos\phi_I)\; d\phi_I}{\left(x_R^2+z_R^2+1-2x_R\cos\phi_I\right)^{3/2}} \end{array}\right\}$$

where our variable of integration is ϕ_I. We cannot replace ϕ_I with ϕ because $\phi \equiv \phi_p$. The variables ϕ_I and ϕ_p are two very different things, so be careful.

Unfortunately, B_x and B_z require numerical integration. However, we can evaluate B_y using a change of variable (something the mathematicians like to call a u-substitution). Choosing how to define the new variable is a bit of an art, but the desired result is always the same: make the integrand as simple as possible. This is done by choosing a definition for the new variable that is as complex as possible such that all forms of the old variable can vanish. In this case,

$$u = x_R^2 + z_R^2 + 1 - 2x_R\cos\phi_I$$

would be the best choice. The first derivative of this is

$$\frac{du}{d\phi_I} = 2x_R\sin\phi_I \quad \Rightarrow \frac{du}{2x_R} = \sin\phi_I\, d\phi_I.$$

This results in a y-component of

$$B_y = \frac{k_M I}{R}\frac{z_R}{2x_R}\int_{u_1}^{u_2} u^{-3/2}du.$$

Using our change of variable, it turns out $u_1 = u_2 = x_R^2 + z_R^2 + 1 - 2x_R$ because $\cos(0) = \cos(2\pi) = 1$, which means $B_y = 0$.

We need to keep in mind here the y-component is only zero because we suppressed the y-component of the position of our arbitrary point p. As mentioned in step 2, this system has cylindrical symmetry about the z-axis. This means the x-component of $\vec{r}_p$ could just as easily measure the distance from the z-axis in any direction parallel to the xy-plane (i.e. x_R is now s_R). These are cylindrical coordinates as described in Section 1.2. In the xz-plane, the y-direction is equivalent to the ϕ-direction. We now have

$$\left\{\begin{array}{rcl} B_s &=& \dfrac{k_M I}{R}\displaystyle\int_0^{2\pi} \frac{z_R\cos\phi_I\, d\phi_I}{\left(s_R^2 + z_R^2 + 1 - 2s_R\cos\phi_I\right)^{3/2}} \\ B_\phi &=& 0 \\ B_z &=& \dfrac{k_M I}{R}\displaystyle\int_0^{2\pi} \frac{(1 - s_R\cos\phi_I)\, d\phi_I}{\left(s_R^2 + z_R^2 + 1 - 2s_R\cos\phi_I\right)^{3/2}} \end{array}\right\} \tag{5.2.11}$$

where $\vec{B} = B_s\hat{s} + B_\phi\hat{\phi} + E_z\hat{z}$. This represents the completely general solution under the generalized coordinates (s, ϕ, z).

Example 5.2.3

A Helmholtz coil is constructed of two circular coils of radius R separated by a distance R, each have N loops of wire. The magnetic field it produces is extremely uniform in between the two coils. To justify this statement, show that a separation of R results in the most uniform field and sketch the magnetic field.

- We'll start with Eq. 5.2.11 to save some time. We can set the point p along the z-axis for simplicity since it will be sufficient to show the uniformity along the axis. Under this assumption, the s-component is

$$B_s = \frac{k_M I}{R}\int_0^{2\pi} \frac{z_R\cos\phi_I\, d\phi_I}{\left(z_R^2 + 1\right)^{3/2}}$$

$$B_s = \frac{k_M I}{R} \frac{z_R}{\left(z_R^2+1\right)^{3/2}} \int_0^{2\pi} \cos\phi_I \, d\phi_I = 0.$$

Therefore, the B-field only has a z-component along the z-axis (i.e. $\vec{B} = B_z \hat{z}$). The field is now

$$\vec{B} = \frac{k_M I}{R} \int_0^{2\pi} \frac{d\phi_I}{\left(z_R^2+1\right)^{3/2}} \hat{z}$$

$$\vec{B} = \frac{k_M I}{R\left(z_R^2+1\right)^{3/2}} \hat{z} \int_0^{2\pi} d\phi_I$$

$$\vec{B} = \frac{2\pi k_M I}{R\left(z_R^2+1\right)^{3/2}} \hat{z}$$

- A coil is simply like having N loops in one place. Therefore, the field is

$$\vec{B} = \frac{2\pi k_M N I}{R\left(z_R^2+1\right)^{3/2}} \hat{z}$$

- This, however, is only generated by a single coil centered at the origin and we have two coils in two different locations. If we shift them each by a from the origin in opposite directions, then the coordinate transformations are $z_{R,bott} = z + a$ for the bottom coil (origin is above it) and $z_{R,top} = z - a$ for the top coil (origin is below it). The quantity z is the location of the arbitrary point p in the new coordinate system as shown in Figure 5.8. This results in a total field of

$$\vec{B} = \vec{B}_{bott} + \vec{B}_{top}$$

$$\vec{B} = \frac{2\pi k_M N I}{R\left[(z+a)^2+1\right]^{3/2}} \hat{z} + \frac{2\pi k_M N I}{R\left[(z-a)^2+1\right]^{3/2}} \hat{z}$$

$$\boxed{\vec{B} = \frac{2\pi k_M N I}{R} \left[\frac{1}{\left[(z+a)^2+1\right]^{3/2}} + \frac{1}{\left[(z-a)^2+1\right]^{3/2}} \right] \hat{z}.}$$

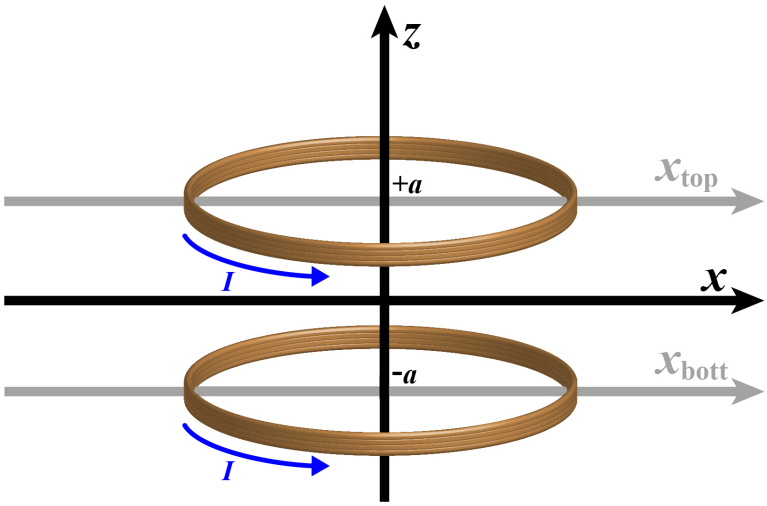


Figure 5.8: This is a two-coil system in which the coils of radius R are separated by a distance of $2a$. If $2a = R$, then this system is called a Helmholtz coil. The coordinate system used for Eq. 5.2.11 is also shown for each individual coil.

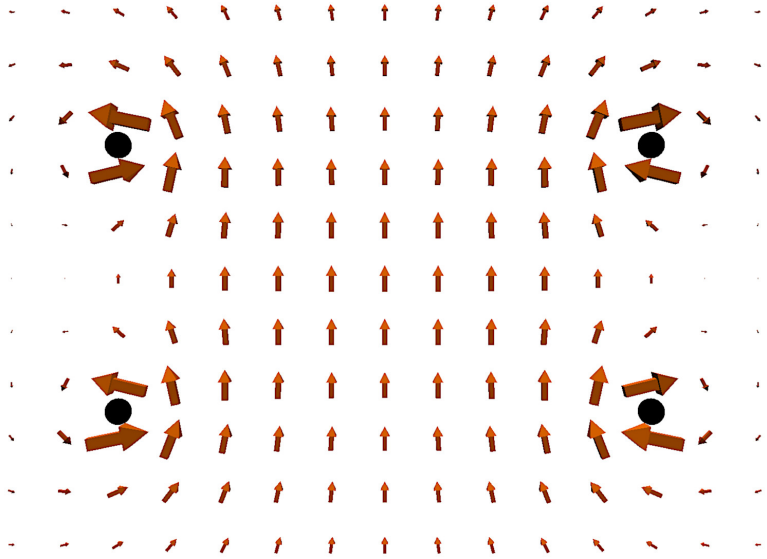


Figure 5.9: This is the magnetic field of the Helmholtz coil (at least in the xz-plane). The large dots are cross-sections of the coils and field strength is indicated by the thickness of the arrows.

This is the magnetic field of a Helmholtz coil at any point along the z-axis given the coils are separated by $2a$. Remember, both z and a are unitless because they're defined in terms of $z_R = z_p/R$.

- To show uniformity, we first need to know how the field is changing along the z-axis. That is given by

$$\frac{d\vec{B}}{dz} = \frac{2\pi k_M NI}{R}\left[\frac{-3\left(z+a\right)}{\left[\left(z+a\right)^2+1\right]^{5/2}} + \frac{-3\left(z-a\right)}{\left[\left(z-a\right)^2+1\right]^{5/2}}\right]\hat{z}$$

$$\frac{d\vec{B}}{dz} = \frac{-6\pi k_M NI}{R}\left[\frac{z+a}{\left[\left(z+a\right)^2+1\right]^{5/2}} + \frac{z-a}{\left[\left(z-a\right)^2+1\right]^{5/2}}\right]\hat{z}.$$

However, this doesn't tell us anything about uniformity. For that, we need to know how the changes are changing, meaning we need a second derivative. The result is

$$\frac{d^2\vec{B}}{dz^2} = \frac{-6\pi k_M NI}{R}\left[\frac{1}{\left[\left(z+a\right)^2+1\right]^{5/2}} + \frac{1}{\left[\left(z-a\right)^2+1\right]^{5/2}} + \frac{-5\left(z+a\right)^2}{\left[\left(z+a\right)^2+1\right]^{7/2}} + \frac{-5\left(z-a\right)^2}{\left[\left(z-a\right)^2+1\right]^{7/2}}\right]\hat{z}.$$

If we want uniformity, then we want the change in $\vec{B}$ to be minimum. This occurs when the second derivative is equal to zero. Furthermore, we want to know this between the coils. If we conveniently chose the origin (the best place between the coils), then

$$0 = \left.\frac{d^2\vec{B}}{dz^2}\right|_{z=0} = \frac{-6\pi k_M NI}{R}\left[\frac{2}{\left(a^2+1\right)^{5/2}} + \frac{-10a^2}{\left(a^2+1\right)^{7/2}}\right]\hat{z}$$

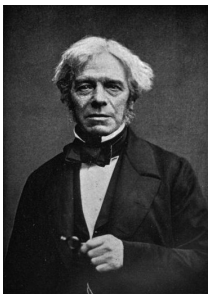
$$0 = \frac{1}{\left(a^2+1\right)^{5/2}} + \frac{-5a^2}{\left(a^2+1\right)^{7/2}}.$$

Now we can multiply through by $\left(a^2+1\right)^{7/2}$ to eliminate the fractions. We now have

$$0 = \left(a^2+1\right) - 5a^2 = -4a^2 + 1 \quad \Rightarrow \boxed{2a = 1}.$$


Hans Ørsted


André-Marie Ampére


Michael Faraday


Carl Gauss

Figure 5.10: These people were important in the development of theoretical electrodynamics.

Recall, that $2a$ was the coil separation in terms of R (i.e. multiples of R). Therefore, the coil separation resulting in the least change in $\vec{B}$ at the center is R. The overall result of this is the extremely uniform field seen in Figure 5.9.

5.3 Theoretical Laws

Eqs. 5.2.6 and 5.2.9 describe the concepts of electric and magnetic fields, respectively. These laws are purely experimental indicating a clear relationship between quantities. However, they do not allow us to understand why the relationships are the way they are nor do they allow us to form conclusions beyond those relationships. This is something which requires a fundamental theoretical understanding of the behavior of E-fields and B-fields.

Ampére's Law

Our theoretical understanding begins with André-Marie Ampére in 1820. Yes, that's the same year Biot and Savart released their findings. Both the Biot-Savart team and Ampére were inspired by Hans Christian Ørsted's discovery that a compass needle pointed perpendicular to a current carrying wire. Ørsted had announced his work in April 1820 and one week later Ampére demonstrated that parallel currents attract and anti-parallel currents

repel. Biot and Savart's work wasn't published until October of that year, so Ampére was already showing promise.

Six years later, Ampére published a memoir in which he presented all his theory and experimental results on magnetism. Amongst other things, it included a beautifully simple relationship between current and B-field we now write as

$$\oint \vec{B} \bullet d\vec{\ell} = \mu_0 I_{enc}\,, \tag{5.3.1}$$

where I_{enc} is the current passing through (i.e. enclosed by) the curve ℓ and μ_0 is a theoretical constant with a value of $4\pi k_M = 4\pi \times 10^{-7}$ N/A^2. Redefining the magnetic constant now makes several results in this chapter look much more elegant. We call Eq. 5.3.1 **Ampére's law**. The closed loop given by the integral is called an **Ampérian loop** and is arbitrarily chosen very much like a coordinate system. Eq. 5.3.1 states that, if there is an electric current inside a closed curve, then there is a magnetic field along that curve. Essentially, moving charge generates a magnetic field (a concept we've already seen).

In an introductory physics textbook, you might see Eq. 5.3.1 used to find the magnetic field generated by an infinitely long current carrying wire or an infinitely long solenoid, but this drastically devalues the law. First, we may be able to find a scenario that approximates one of these possibilities, but neither truly exists. Second, other than these few rare occurrences, the Biot-Savart law is far more practical for finding a B-field. Ampére's law can be used to find an electric current given a magnetic field, but it has a higher purpose. It gives us a much better understanding of how magnetic fields work, the depth of which was not seen clearly until years later (the majority of the scientific community initially favored the Biot-Savart law).

To get a feel for the real theoretical power of Ampére's law, we need to use something called the Curl Theorem given by Eq. 3.5.12. With it, we can write Eq. 5.3.1 as

$$\int \left(\vec{\nabla} \times \vec{B}\right) \bullet d\vec{A} = \mu_0 I_{enc}$$

where $\vec{\nabla}$ is the del operator (defined in Chapter 3). We can simplify this by defining a current density (current per unit area) with

$$I = \int \vec{J} \bullet d\vec{A}, \tag{5.3.2}$$

where $\vec{J}$ is the current density and I is the current. If we integrate the current density over the same area as the one enclosed by the Ampérian loop, then I becomes I_{enc} and we have

$$\int \left(\vec{\nabla} \times \vec{B}\right) \bullet d\vec{A} = \mu_0 \int \vec{J} \bullet d\vec{A}.$$

$$\int \left(\vec{\nabla} \times \vec{B}\right) \bullet d\vec{A} = \int \mu_0 \vec{J} \bullet d\vec{A}.$$

Since the areas of integration are the same, we can just cancel them (using Eq. 3.1.1) leaving us with

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{J}, \tag{5.3.3}$$

which is defined at a single arbitrary point. Eq. 5.3.3 tells us the curl of the magnetic field at a point in space is directly proportional to the current density at that same point. This is a very powerful idea because it relates magnetic fields and current in terms of vector calculus as described in Chapter 3.

Example 5.3.1

Show that the Biot-Savart law is consistent with Ampére's law.

- First, we need to make the Biot-Savart law look a little more convenient. We'll start with the current density form which is given by

$$\vec{B} = k_M \int \frac{\vec{J} \times \hat{r}}{r^2} dV_I = k_M \int \vec{J} \times \frac{\vec{r}}{r^3} dV_I$$

where we have eliminated the unit vector using Eq. 5.2.4. Generalizing further, we get

$$\vec{B} = k_M \int \vec{J}(\vec{r_I}) \times \frac{(\vec{r_p} - \vec{r_I})}{|\vec{r_p} - \vec{r_I}|^3} dV_I \tag{5.3.4}$$

where we have taken extra care in showing $\vec{J}$ is only dependent on the position of the current and not the position of the arbitrary point p.

- Now we're going to make a very creative substitution using the del operator. Let take

$$\vec{\nabla}_p\left(\frac{1}{r}\right) = \vec{\nabla}_p\left(\frac{1}{|\vec{r}_p - \vec{r}_I|}\right)$$

where the subscript of p on del indicates the derivatives are with respect to $\vec{r}_p$, not $\vec{r}$. It's best to evaluate this gradient in Cartesian coordinates, yet the result will hold for any coordinate system. We'll be using Eq. 3.2.1 and

$$|\vec{r}_p - \vec{r}_I| = \sqrt{(x_p - x_I)^2 + (y_p - y_I)^2 + (y_p - y_I)^2}$$

to get components of

$$\left\{\begin{array}{rcl} \hat{x}\dfrac{\partial}{\partial x_p}\left(\dfrac{1}{r}\right) & = & \dfrac{-(x_p - x_I)\,\hat{x}}{\left[(x_p - x_I)^2 + (y_p - y_I)^2 + (y_p - y_I)^2\right]^{3/2}} \\ \hat{y}\dfrac{\partial}{\partial y_p}\left(\dfrac{1}{r}\right) & = & \dfrac{-(y_p - y_I)\,\hat{y}}{\left[(x_p - x_I)^2 + (y_p - y_I)^2 + (y_p - y_I)^2\right]^{3/2}} \\ \hat{z}\dfrac{\partial}{\partial z_p}\left(\dfrac{1}{r}\right) & = & \dfrac{-(z_p - z_I)\,\hat{z}}{\left[(x_p - x_I)^2 + (y_p - y_I)^2 + (y_p - y_I)^2\right]^{3/2}} \end{array}\right\}$$

and a total result of

$$\vec{\nabla}_p\left(\frac{1}{r}\right) = \frac{-(x_p - x_I)\,\hat{x} - (y_p - y_I)\,\hat{y} - (z_p - z_I)\,\hat{z}}{\left[(x_p - x_I)^2 + (y_p - y_I)^2 + (y_p - y_I)^2\right]^{3/2}}$$

$$\boxed{\vec{\nabla}_p\left(\frac{1}{r}\right) = -\frac{\vec{r}}{r^3}} \tag{5.3.5}$$

where $\vec{r} = \vec{r}_p - \vec{r}_I$.

- If we substitute Eq. 5.3.5 into Eq. 5.3.4, we get

$$\vec{B} = k_M \int \vec{J}(\vec{r}_I) \times -\vec{\nabla}_p\left(\frac{1}{|\vec{r}_p - \vec{r}_I|}\right) dV_I$$

$$\vec{B} = k_M \int -\vec{J}(\vec{r_I}) \times \vec{\nabla}_p \left(\frac{1}{|\vec{r}_p - \vec{r}_I|} \right) dV_I.$$

If we use the derivative product rule given by Eq. 3.2.11 (the first term on the right is our integrand), then the result is

$$\vec{B} = k_M \int \left[\vec{\nabla}_p \times \frac{\vec{J}(\vec{r_I})}{|\vec{r}_p - \vec{r}_I|} - \frac{1}{|\vec{r}_p - \vec{r}_I|} \left(\vec{\nabla}_p \times \vec{J}(\vec{r_I}) \right) \right] dV_I.$$

Since $\vec{J}(\vec{r_I})$ is not dependent on $\vec{r}_p$, the second term in the integrand is zero. This leaves us with

$$\vec{B} = k_M \int \vec{\nabla}_p \times \frac{\vec{J}(\vec{r_I})}{|\vec{r}_p - \vec{r}_I|} dV_I.$$

We can now pull the curl with respect to $\vec{r}_p$ out of the integral entirely because the integral is with respect to $\vec{r}_I$. Therefore,

$$\vec{B} = \vec{\nabla}_p \times \left[k_M \int \frac{\vec{J}(\vec{r_I})}{|\vec{r}_p - \vec{r}_I|} dV_I \right]. \tag{5.3.6}$$

It's good to note here that the quantity in square brackets is the magnetic vector potential (something we'll get into a little later in the chapter). At this point, you might be thinking "Will this solution ever end?!" I assure you, in being this thorough, the following examples will be incredibly simple in comparison. It's important that we get all this out of the way.

- Ampére's law given by Eq. 5.3.3 involves the curl of $\vec{B}$, so

$$\vec{\nabla}_p \times \vec{B} = \vec{\nabla}_p \times \left(\vec{\nabla}_p \times \left[k_M \int \frac{\vec{J}(\vec{r_I})}{|\vec{r}_p - \vec{r}_I|} dV_I \right] \right).$$

Using the second derivative identity given by Eq. 3.2.8, we get

$$\vec{\nabla}_p \times \vec{B} = \vec{\nabla}_p \left(\vec{\nabla}_p \bullet \left[k_M \int \frac{\vec{J}(\vec{r_I})}{|\vec{r}_p - \vec{r}_I|} dV_I \right] \right) - \vec{\nabla}_p^2 \left[k_M \int \frac{\vec{J}(\vec{r_I})}{|\vec{r}_p - \vec{r}_I|} dV_I \right].$$

This is looking rather complicated, so let's see what we can do about eliminating some things.

- Let's look at the part of the first term inside the parentheses (call it $\vec{O}$),

$$\vec{O} = \vec{\nabla}_p \bullet \left[k_M \int \frac{\vec{J}(\vec{r_I})}{|\vec{r_p} - \vec{r_I}|} dV_I \right]$$

$$\vec{O} = k_M \int \vec{\nabla}_p \bullet \left[\frac{\vec{J}(\vec{r_I})}{|\vec{r_p} - \vec{r_I}|} \right] dV_I.$$

If we use the derivative product rule given by Eq. 3.2.10 (taking $\vec{\nabla}_p \bullet \vec{J}(\vec{r_I}) = 0$ because $\vec{J}(\vec{r_I})$ is not dependent on $\vec{r_p}$), then the result is

$$\vec{O} = k_M \int \vec{J}(\vec{r_I}) \bullet \vec{\nabla}_p \left(\frac{1}{|\vec{r_p} - \vec{r_I}|} \right) dV_I.$$

It kind of looks like we've just pulled the $\vec{J}(\vec{r_I})$ out of the derivative, but if you look close enough you'll see the divergence changed to a gradient. Don't jump to conclusions too quickly. A similar derivation to the one for Eq. 5.3.5 will give us

$$\vec{\nabla}_p \left(\frac{1}{r} \right) = -\vec{\nabla}_I \left(\frac{1}{r} \right)$$

as a substitution. Using it, we get

$$\vec{O} = k_M \int \vec{J}(\vec{r_I}) \bullet -\vec{\nabla}_I \left(\frac{1}{|\vec{r_p} - \vec{r_I}|} \right) dV_I$$

$$\vec{O} = -k_M \int \vec{J}(\vec{r_I}) \bullet \vec{\nabla}_I \left(\frac{1}{|\vec{r_p} - \vec{r_I}|} \right) dV_I.$$

Now, we'll use Eq. 3.2.10 again (with a little manipulation) to get

$$\vec{O} = -k_M \int \left[\vec{\nabla}_I \bullet \frac{\vec{J}(\vec{r_I})}{|\vec{r_p} - \vec{r_I}|} - \frac{1}{|\vec{r_p} - \vec{r_I}|} \left(\vec{\nabla}_I \bullet \vec{J}(\vec{r_I}) \right) \right] dV_I.$$

The $\vec{\nabla}_I \bullet \vec{J}(\vec{r_I})$ doesn't go to zero as easily as $\vec{\nabla}_p \bullet \vec{J}(\vec{r_I})$ did. However, the Biot-Savart law requires a steady current, which means no charge

can "bunch up" anywhere. Under this approximation, any divergence of $\vec{J}(\vec{r_I})$ must be zero. This leaves us with

$$\vec{O} = -k_M \int \vec{\nabla}_I \bullet \frac{\vec{J}(\vec{r_I})}{|\vec{r_p} - \vec{r_I}|} dV_I.$$

This looks a lot like what we started with, but we needed the del to be with respect to $\vec{r_I}$ before we could perform the next step. If we apply the Divergence Theorem (Eq. 3.5.5), we get

$$\vec{O} = -k_M \oint \frac{\vec{J}(\vec{r_I})}{|\vec{r_p} - \vec{r_I}|} \bullet d\vec{A}_I.$$

What we now have is an integral over a closed surface which is a little easier to understand. The surface in question is the one completely enclosing the volume from the Biot-Savart law (Eq. 5.3.4). This volume is defined such that it includes *all* the current. Therefore, there is no current passing through the surface and we get $\vec{O} = 0$.

- All this leaves us with

$$\vec{\nabla}_p \times \vec{B} = -\vec{\nabla}_p^2 \left[k_M \int \frac{\vec{J}(\vec{r_I})}{|\vec{r_p} - \vec{r_I}|} dV_I \right]$$

and again, $\vec{J}(\vec{r_I})$ is not dependent on $\vec{r}_p$, so

$$\vec{\nabla}_p \times \vec{B} = k_M \int \vec{J}(\vec{r_I}) \left[-\vec{\nabla}_p^2 \left(\frac{1}{|\vec{r_p} - \vec{r_I}|} \right) \right] dV_I.$$

We're almost done!

- Now we need another substitution involving a del, but this one will take a little more thought. Using Eq. 5.3.5, we get

$$-\vec{\nabla}_p^2 \left(\frac{1}{r} \right) = \vec{\nabla}_p \bullet \left[-\vec{\nabla}_p \left(\frac{1}{r} \right) \right]$$

$$-\vec{\nabla}_p^2 \left(\frac{1}{r} \right) = \vec{\nabla}_p \bullet \left(\frac{\vec{r}}{r^3} \right)$$

Just so this doesn't get too messy, we're going to assume $\vec{r}_I$ is zero meaning $\vec{r} = \vec{r}_p$ (don't worry, we'll put it back in later). Now, we have

$$\vec{\nabla}_p \bullet \left(\frac{\vec{r_p}}{r_p^3} \right) = \vec{\nabla}_p \bullet \left(\frac{\hat{r_p}}{r_p^2} \right).$$

In spherical coordinates, we can use Eq. 3.3.7 to arrive at

$$\vec{\nabla}_p \bullet \left(\frac{\vec{r_p}}{r_p^3} \right) = \frac{1}{r_p^2} \frac{\partial}{\partial r_p} \left(r_p^2 \frac{1}{r_p^2} \right) = \frac{1}{r_p^2} \frac{\partial}{\partial r_p} (1) = 0.$$

However, the Divergence Theorem (Eq. 3.5.5) tells us that

$$\int \vec{\nabla}_p \bullet \left(\frac{\hat{r_p}}{r_p^2} \right) dV = \oint \left(\frac{\hat{r_p}}{r_p^2} \right) \bullet d\vec{A}$$

where both integrals enclose the origin (i.e. $\vec{r}_I$ for our purposes). Since the volume (and the surface enclosing) it are arbitrary, we'll choose a sphere of radius a. The line integral on the right gives

$$\oint \left(\frac{\hat{r_p}}{a^2} \right) \bullet \left(a^2 \sin\theta \, d\theta \, d\phi \, \hat{r_p} \right) = \oint \sin\theta \, d\theta \, d\phi = 4\pi$$

which is most definitely not zero. The discrepancy comes from the origin, our $\vec{r}_I$. The divergence goes to infinity at this location, but is zero everywhere else. There is only one entity that has an infinite value at one place, a zero value everywhere else, and also has a finite area underneath: **the Dirac delta function**. Calling it a "function" is misleading since a function must have a finite value everywhere by definition, but the name suffices. The area under this function is 1, but the area under our function is 4π. Therefore, we can conclude that

$$\vec{\nabla}_p \bullet \left(\frac{\vec{r_p}}{r_p^3} \right) = 4\pi \delta^3(\vec{r_p}).$$

where the cube on the δ indicates we're working in 3 dimensions (i.e. $\delta^3(\vec{r}) \equiv \delta(x)\,\delta(y)\,\delta(z)$). We can now put the shift of $\vec{r}_I$ back in since $\vec{r} = \vec{r}_p - \vec{r}_I$ and we get

$$\vec{\nabla}_p \bullet \left(\frac{\vec{r}}{r^3} \right) = 4\pi \delta^3(\vec{r})$$

or even better for us

$$-\vec{\nabla}_p^2\left(\frac{1}{r}\right) = \vec{\nabla}_p \bullet \left(\frac{\vec{r}}{r^3}\right) = 4\pi\delta^3(\vec{r})\,. \tag{5.3.7}$$

- Now the curl of $\vec{B}$ is

$$\vec{\nabla}_p \times \vec{B} = k_M \int \vec{J}(\vec{r_I})\, 4\pi\delta^3(\vec{r_p} - \vec{r_I})\, dV_I$$

$$\vec{\nabla}_p \times \vec{B} = 4\pi k_M \int \vec{J}(\vec{r_I})\, \delta^3(\vec{r_p} - \vec{r_I})\, dV_I.$$

Inside an integral, the Dirac delta function "picks out" where it is non-zero for all other functions in the integrand. For our integral, this would be

$$\vec{\nabla}_p \times \vec{B} = 4\pi k_M \vec{J}(\vec{r_p}) \int \delta^3(\vec{r_p} - \vec{r_I})\, dV_I.$$

The integral now has a value of 1 and $4\pi k_M = \mu_0$, so we get

$$\vec{\nabla}_p \times \vec{B} = \mu_0 \vec{J}(\vec{r_p})\,,$$

which is exactly Eq. 5.3.3.

Faraday's Law

A British scientist by the name of Michael Faraday had been conducting some experiments involving electric current and magnetic fields in the 1820s. He was not formally educated, having learned science while reading books during a seven-year apprenticeship at a book store in his early twenties. This makes the set of contributions he made to science (e.g. the electric motor) in his lifetime very impressive. In 1831, Faraday announced his results regarding how changing magnetic fields could affect electric current. With his limited math skills, the relationship he published was very basic in terms of the application to which he thought it applied.

However, the scope of his relationship was very quickly realized by other scientists who took it upon themselves to generalize the result to

$$\oint \vec{E} \bullet d\vec{\ell} = -\frac{\partial \Phi_B}{\partial t}, \tag{5.3.8}$$

which we call **Faraday's law**. The quantity being differentiated on the right is

$$\Phi_B = \int \vec{B} \bullet d\vec{A}, \tag{5.3.9}$$

which we call the **magnetic flux.** It is called flux because its form is analogous to flux from fluid dynamics,

$$\Phi_{\text{fluid}} = \int \rho \vec{v} \bullet d\vec{A}, \tag{5.3.10}$$

where ρ is the fluid density and $\vec{v}$ is the flow velocity through the area of integration. In reality, magnetic fields don't flow, but vector fields can still be discussed in flow terms even if there isn't anything flowing as long as there is a non-zero curl. The curl of the magnetic field is given by Eq. 5.3.3, which is non-zero (at some points).

Eq. 5.3.8 states that, if a magnetic field changes on some area, then there is an electric field along the curve enclosing that area. Essentially, a changing magnetic field generates an electric field. This idea has much more broad a scope than Michael Faraday had anticipated. It forms the foundation for AC circuit designs and led the great Nikola Tesla (for which the standard unit of magnetic field is named) to the design the entire U.S. electricity grid at the turn of the 20th century.

Just as with Ampére's law (Eq. 5.3.1), we have a line integral on the left, so we can get a feel for its theoretical power by applying the Curl Theorem (Eq. 3.5.12). Doing so, we arrive at

$$\int \left(\vec{\nabla} \times \vec{E}\right) \bullet d\vec{A} = -\frac{\partial \Phi_B}{\partial t}.$$

Substituting in for magnetic flux with Eq. 5.3.9, we get

$$\int \left(\vec{\nabla} \times \vec{E}\right) \bullet d\vec{A} = -\frac{\partial}{\partial t} \int \vec{B} \bullet d\vec{A}.$$

The integral operator is over space and the derivative operator is over time, so these operators are commutative. Applying this property results in

$$\int \left(\vec{\nabla} \times \vec{E}\right) \bullet d\vec{A} = \int -\frac{\partial \vec{B}}{\partial t} \bullet d\vec{A}.$$

Since the areas of integration are the same, we can just cancel them (using Eq. 3.1.1) leaving us with

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}, \tag{5.3.11}$$

which is defined at a single arbitrary point. Eq. 5.3.11 tells us the curl of the electric field at a point in space is directly proportional to the rate of change of the magnetic field with respect to time at that same point. This is a very powerful idea because it relates electric fields and magnetic fields in terms of vector calculus as described in Chapter 3.

Example 5.3.2

Show that Coulomb's law is consistent with Faraday's law.

- First, we need to make Coulomb's law look a little more convenient. Starting with Eq. 5.2.7, we can generalize using $dq = \rho\, dV$ to get

$$\vec{E} = k_E \int \frac{\rho\, dV}{|\vec{r}_p - \vec{r}_q|^3} \left(\vec{r}_p - \vec{r}_q\right).$$

  Now we'll be a little more specific with dependencies. The integral is over the volume of charge and the charge density is only dependent on the location of the charge, so

$$\vec{E} = k_E \int \rho(\vec{r}_q) \frac{(\vec{r}_p - \vec{r}_q)}{|\vec{r}_p - \vec{r}_q|^3}\, dV_q. \tag{5.3.12}$$

- Faraday's law given by Eq. 5.3.11 involves the curl of $\vec{E}$, so

$$\vec{\nabla}_p \times \vec{E} = \vec{\nabla}_p \times \left[k_E \int \rho(\vec{r}_q) \frac{(\vec{r}_p - \vec{r}_q)}{|\vec{r}_p - \vec{r}_q|^3}\, dV_q\right].$$

Since both the variable of integration and $\rho(\vec{r}_q)$ are independent on $\vec{r}_p$, we get

$$\vec{\nabla}_p \times \vec{E} = k_E \int \rho(\vec{r}_q)\, \vec{\nabla}_p \times \left[\frac{(\vec{r}_p - \vec{r}_q)}{|\vec{r}_p - \vec{r}_q|^3} \right] dV_q.$$

Making a substitution from Eq. 5.3.7, this becomes

$$\vec{\nabla}_p \times \vec{E} = k_E \int \rho(\vec{r}_q)\, \vec{\nabla}_p \times \left[-\vec{\nabla}_p \left(\frac{1}{|\vec{r}_p - \vec{r}_q|} \right) \right] dV_q$$

$$\vec{\nabla}_p \times \vec{E} = -k_E \int \rho(\vec{r}_q)\, \vec{\nabla}_p \times \left[\vec{\nabla}_p \left(\frac{1}{|\vec{r}_p - \vec{r}_q|} \right) \right] dV_q.$$

Since the curl of a gradient is always zero (Eq. 3.2.6), the integrand is zero. Therefore,

$$\vec{\nabla}_p \times \vec{E} = 0,$$

which is Faraday's law given that $\vec{B}$ doesn't change in time (something true even for the Biot-Savart law).

Gauss's Law(s)

The next major discovery came in 1835 with Carl Friedrich Gauss, a German mathematician and scientist. Gauss formulated relationships for electricity and magnetism in terms of flux through closed areas. They are formally written today as

$$\oint \vec{E} \bullet d\vec{A} = \frac{q_{enc}}{\epsilon_0} \tag{5.3.13}$$

and

$$\oint \vec{B} \bullet d\vec{A} = 0\,, \tag{5.3.14}$$

where q_{enc} is the charge enclosed by the area given in the closed surface integral and ϵ_0 is a theoretical constant with a value of $(4\pi k_E)^{-1} = 8.854 \times 10^{-12}$ C^2/(Nm^2). Redefining the electric constant now makes several results in this chapter look much more elegant. We call Eq. 5.3.13 **Gauss's law.** Eq. 5.3.14 doesn't have a formal name, but we sometimes call it **Gauss's law for Magnetism**. The closed area given by the integrals is called a **Gaussian Surface** and is arbitrarily chosen very much like a coordinate system.

Eq. 5.3.13 states that, if there is an electric charge inside a closed surface, then there is a net electric field passing through that surface (i.e. an electric flux through the surface as analogous to Eq. 5.3.10). Essentially, charge generates an electric field (a concept we've already seen). Eq. 5.3.13 states that there isn't a magnetic flux through *any* closed surface because the integral is necessarily zero. No matter what shape, size, orientation, or location this arbitrary surface has, there are always as many vectors on the surface directed inward as there are directed outward. Essentially, this means magnetic fields always form closed loops (i.e. they always lead back to the source).

Because the integrals in Eqs. 5.3.13 and 5.3.14 are both closed surface integrals, we can apply something called the Divergence Theorem (Eq. 3.5.5) to get a feel for their theoretical power. Showing the work for Eq. 5.3.13, we see that

$$\int \vec{\nabla} \bullet \vec{E}\, dV = \frac{q_{enc}}{\epsilon_0}.$$

We can simplify this by defining a charge density (charge per unit volume) with

$$q = \int \rho\, dV, \tag{5.3.15}$$

where ρ is the charge density and q is the charge. If we integrate the charge density over the same volume as the one enclosed by the Gaussian Surface, then q becomes q_{enc} and we have

$$\int \vec{\nabla} \bullet \vec{E}\, dV = \frac{1}{\epsilon_0} \int \rho\, dV$$

$$\int \vec{\nabla} \bullet \vec{E}\, dV = \int \frac{\rho}{\epsilon_0}\, dV.$$

Since the volumes of integration are the same, we can just cancel them (using Eq. 3.1.1) leaving us with

$$\vec{\nabla} \bullet \vec{E} = \frac{\rho}{\epsilon_0}, \tag{5.3.16}$$

which is defined at a single arbitrary point. Eq. 5.3.16 tells us the divergence of the electric field at a point in space is directly proportional to the charge density at that same point. This is a very powerful idea because it relates electric fields and charge in terms of vector calculus as described in Chapter 3.

Similarly, Eq. 5.3.14 can be shown to become

$$\vec{\nabla} \bullet \vec{B} = 0, \tag{5.3.17}$$

which is defined at a single arbitrary point. Eq. 5.3.17 tells us the divergence of the magnetic field at *any* point in space is zero (i.e. magnetic fields don't diverge). This is a very powerful idea because it shows the behavior of magnetic fields in terms of vector calculus as described in Chapter 3.

Example 5.3.3

Show that Coulomb's law is consistent with Gauss's law.

- Starting with Eq. 5.3.12 and taking it's divergence, we have

$$\vec{\nabla}_p \bullet \vec{E} = \vec{\nabla}_p \bullet \left[k_E \int \rho(\vec{r}_q) \frac{(\vec{r}_p - \vec{r}_q)}{|\vec{r}_p - \vec{r}_q|^3} \, dV_q \right].$$

  Since both the variable of integration and $\rho(\vec{r}_q)$ are independent on $\vec{r}_p$, we get

$$\vec{\nabla}_p \bullet \vec{E} = k_E \int \rho(\vec{r}_q) \, \vec{\nabla}_p \bullet \left[\frac{(\vec{r}_p - \vec{r}_q)}{|\vec{r}_p - \vec{r}_q|^3} \right] dV_q.$$

  By Eq. 5.3.7, this integral simplifies to

$$\vec{\nabla}_p \bullet \vec{E} = k_E \int \rho(\vec{r}_q) \, 4\pi \delta^3(\vec{r}_p - \vec{r}_q) \, dV_q.$$

Inside an integral, the Dirac delta function "picks out" where it is non-zero for all other functions in the integrand. For our integral, this would be

$$\vec{\nabla}_p \bullet \vec{E} = 4\pi k_E \, \rho(\vec{r}_p) \int \delta^3(\vec{r}_p - \vec{r}_q) \, dV_q.$$

The integral now has a value of 1 and $\epsilon_0 = (4\pi k_E)^{-1}$, so we get

$$\vec{\nabla}_p \bullet \vec{E} = \frac{\rho(\vec{r}_p)}{\epsilon_0},$$

which is exactly Eq. 5.3.16.

Example 5.3.4

Show that the Biot-Savart law is consistent with Gauss's law for Magnetism.

- Starting with Eq. 5.3.6 and taking it's divergence, we have

$$\vec{\nabla}_p \bullet \vec{B} = \vec{\nabla}_p \bullet \left(\vec{\nabla}_p \times \left[k_M \int \frac{\vec{J}(\vec{r_I})}{|\vec{r}_p - \vec{r}_I|} dV_I \right] \right). \tag{5.3.18}$$

  Since the divergence of a curl is zero (Eq. 3.2.7), this results in

$$\vec{\nabla}_p \bullet \vec{B} = 0, \tag{5.3.19}$$

  which is exactly Eq. 5.3.17.

Ampére's Law Revisted

Almost 30 years after Gauss, a Scottish physicist named James Clerk Maxwell was pondering Ampére's law, given by Eq. 5.3.1, and felt there was something missing. At this point, Maxwell is still under the presumption that current is a flowing fluid because we're not even sure atoms exist, let alone charged particles like electrons. Maxwell envisions a vortex sea within the fluid inside

his dielectric materials responding to the presence of external fields. These vorticies represent an extra form of motion for the fluid and, therefore, should require an extra electric current term in Eq. 5.3.1.

In 1861, Maxwell published a paper called *On Physical Lines of Force* where he laid out a new Ampére's law given by

$$\oint \vec{B} \bullet d\vec{\ell} = \mu_0 I_{enc} + \mu_0 I_D,$$

where I_D is the **displacement current** representing the extra displacement in the electric fluid (that doesn't really exist). We can use the Curl Theorem (Eq. 3.5.12) just as we did for Ampére's law in Section 5.3 to arrive at

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{J} + \mu_0 \vec{J}_D. \tag{5.3.20}$$

Another way to think about this is to tap another fluid dynamics concept: **equations of continuity**. The basic fluid form of this would be

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \bullet (\rho \vec{v}) = 0, \tag{5.3.21}$$

which is very related to the fluid flux given by Eq. 5.3.10. Formulating this for electrodynamics, we get

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \bullet \vec{J} = 0,$$

or

$$\vec{\nabla} \bullet \vec{J} = -\frac{\partial \rho}{\partial t}, \tag{5.3.22}$$

where ρ is the volumetric charge density and $\vec{J}$ is the current density (current per unit area). This is commonly referred to as **Conservation of Charge** because it states the spatial flow of charge (current density) outward from a point in space is equal to the decrease in the charge density over time at that same point. Seems logical, right? We can see, using vector calculus, Ampére's law given by Eq. 5.3.3 is not consistent with Eq. 5.3.22.

According to Eq. 3.2.7, we know the divergence of a curl is always zero. If we take the divergence of Eq. 5.3.3, we get

$$\vec{\nabla} \bullet \left(\vec{\nabla} \times \vec{B} \right) = \vec{\nabla} \bullet \left(\mu_0 \vec{J} \right)$$

$$0 = \vec{\nabla} \bullet \vec{J}.$$

This doesn't match Eq. 5.3.22, so there must be something missing from Ampére's law. Working this out in terms of vector calculus allows us to discover the true origin of the displacement current. Taking the divergence of Eq. 5.3.20, we get

$$\vec{\nabla} \bullet \left(\vec{\nabla} \times \vec{B}\right) = \vec{\nabla} \bullet \left(\mu_0 \vec{J}\right) + \vec{\nabla} \bullet \left(\mu_0 \vec{J}_D\right)$$

$$0 = \vec{\nabla} \bullet \vec{J} + \vec{\nabla} \bullet \vec{J}_D.$$

Because of Eq. 5.3.22, this implies that

$$\vec{\nabla} \bullet \vec{J}_D = \frac{\partial \rho}{\partial t}.$$

From Gauss's law given by Eq. 5.3.16, we can say

$$\vec{\nabla} \bullet \vec{J}_D = \frac{\partial}{\partial t}\left(\epsilon_0 \vec{\nabla} \bullet \vec{E}\right).$$

The del operator is over space, so it is commutative with the time derivative. Applying this property results in

$$\vec{\nabla} \bullet \vec{J}_D = \vec{\nabla} \bullet \left(\epsilon_0 \frac{\partial \vec{E}}{\partial t}\right)$$

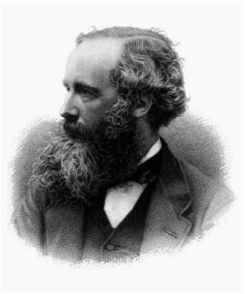
$$\vec{J}_D = \epsilon_0 \frac{\partial \vec{E}}{\partial t}.$$

The so-called displacement current term is simply the result of a changing electric field! We can substitute this result into Eq. 5.3.20 and we get

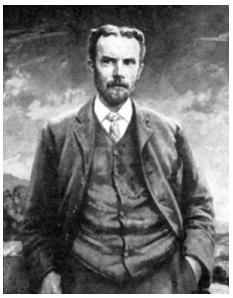
$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{J} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \tag{5.3.23}$$

and an integral form of

$$\oint \vec{B} \bullet d\vec{\ell} = \mu_0 I_{enc} + \mu_0 \epsilon_0 \frac{\partial \Phi_E}{\partial t}, \tag{5.3.24}$$



James Clerk Maxwell



Oliver Heaviside

Figure 5.11: These people were important in the development of what we call Maxwell's equations.

where Φ_E is the electric flux passing through the area enclosed by the curve in the line integral. The integral form of these laws is appealing to some, but we have seen very clearly in the examples from Section 5.3 and the immediately preceding work that the del form is far more powerful. It's also appropriate at this point in our discussion to stick to the del form because Maxwell was the first to formally use the notation.

5.4 Unification of Electricity and Magnetism

Discovery of the displacement current was a major step in the development of electrodynamics. It led Maxwell to another major publication only a few years later. In 1865, Maxwell published a paper called *A Dynamical Theory of the Electromagnetic Field* where he listed many equations together becoming the first to truly unify electricity with magnetism under one theory. The list included 20 equations, but his notation was atrocious. We can compress that to 8 equations using vector notation and using more familiar quantities, symbols, units, and names.

- *"Total Motion of Electricity" (Definition of Total Current)*:

$$\vec{J}_{tot} = \vec{J} + \frac{\partial \vec{D}}{\partial t}, \tag{5.4.1}$$

where $\vec{D}$ is the displacement field (i.e. the electric field in the material).

- *"Equation of Magnetic Intensity" (Definition of $\vec{H}$ and $\vec{A}$)*:

$$\vec{B} = \mu\vec{H} = \vec{\nabla} \times \vec{A}, \tag{5.4.2}$$

where μ is a magnetic field constant for the material, $\vec{H}$ is hysteresis field (i.e. the magnetic field in the material), and $\vec{A}$ is the magnetic vector potential.

- *"Equation of Current" (Ampére's Law for Materials)*:

$$\vec{\nabla} \times \mu\vec{H} = \mu\vec{J}_{tot}, \tag{5.4.3}$$

where μ is a magnetic field constant for the material and $\vec{H}$ is hysteresis field (i.e. the magnetic field in the material). This equation is just Eq. 5.3.23 applied to materials.

- *"Equation of Electromotive Force" (Total Electromagnetic Field)*:

$$\left[\vec{v} \times \vec{B} + \vec{E}\right] = \vec{v} \times \mu\vec{H} - \frac{\partial\vec{A}}{\partial t} - \vec{\nabla}\phi, \tag{5.4.4}$$

where μ is a magnetic field constant for the material, $\vec{H}$ is hysteresis field (i.e. the magnetic field in the material), $\vec{A}$ is the magnetic vector potential, and ϕ is the electric potential.

- *"Equation of Electric Elasticity" (Definition of $\vec{D}$)*:

$$\vec{E} = \frac{1}{\epsilon}\vec{D}, \tag{5.4.5}$$

where ϵ is an electric field constant for the material and $\vec{D}$ is the displacement field (i.e. the electric field in the material).

- *"Equation of Electric Resistance" (Ohm's Law)*:

$$\vec{E} = \frac{1}{\sigma}\vec{J}, \tag{5.4.6}$$

where σ is the electric conductivity in the material.

- *"Equation of Free Electricity" (Gauss's Law for Materials)*:

$$\vec{\nabla} \bullet \vec{D} = \rho, \tag{5.4.7}$$

  where $\vec{D}$ is the displacement field (i.e. the electric field in the material) and ρ is the volumetric charge density in the material.

- *"Equation of Continuity" (Charge Conservation)*:

$$\vec{\nabla} \bullet \vec{J} = -\frac{\partial \rho}{\partial t}, \tag{5.4.8}$$

  where ρ is the volumetric charge density in the material. This is just Eq. 5.3.22.

The quantities $\vec{D}$, $\vec{H}$, ϵ, μ, and σ are all related in some way to materials. Maxwell was experimental at heart, so he designed the equations for practical use rather than deeper meaning. In fact, he viewed the electric potential, ϕ, and the magnetic potential, $\vec{A}$, as completely meaningless because where you chose to place the value of zero was irrelevant. Very much like a coordinate system (see Chapter 1), this choice of zero has no effect on the physical result, but there are some choices that will simplify the analysis. Both ϕ and $\vec{A}$ had been used prior to Maxwell by people like Joseph Louis Lagrange, Pierre-Simon Laplace, Gustav Kirchhoff, Michael Faraday, and Franz Neumann; all of whom tried to interpret them physically to no real success. Maxwell, on the other hand, simply viewed them as a way to simplify his equations.

We could very easily combine several of these equation to simplify the work required and hopefully make the list look a little more elegant. In fact, Oliver Heaviside, an English mathematician and physicist, did just that. Heaviside's major contributions include formalizing the notation we use in vector calculus given in Chapter 3, developing methods of solving differential equations, and incorporating complex numbers into the methods of electric circuits. In 1885, he published *Electromagnetic Induction and its Propagation* where he took Maxwell's list of 8 down to 4 equations.

Heaviside realized, not only could he combine a few of Maxwell's equations to shorten the list, he could eliminate several equations and arbitrarily defined quantities by including Faraday's law (Eq. 5.3.11). He felt that, since Maxwell's arbitrary quantities had no physical meaning, they should not be included. In response, Maxwell spent years trying to discover their physical

significance with absolutely no success and ultimately conceded to Heaviside on the issue in 1868.

Heaviside's list also generalized the equations for use everywhere rather than just in materials and they can be used to derive all of the equations on Maxwell's list. Heaviside brought together the work of Gauss, Faraday, and Ampére under the mathematics of vector calculus to provide us with

$$\vec{\nabla} \bullet \vec{E} = \frac{\rho}{\epsilon_0} \tag{5.4.9a}$$

$$\vec{\nabla} \bullet \vec{B} = 0 \tag{5.4.9b}$$

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{5.4.9c}$$

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{J} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \tag{5.4.9d}$$

which are just Eqs. 5.3.16, 5.3.17, 5.3.11 and 5.3.23. These equations are formulated in terms of just the electric and magnetic fields. Heaviside also listed the **Lorentz Force** as

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B} \tag{5.4.10}$$

to incorporate how charges were affected by each of these fields. In this case, the electric field constant, ϵ_0, is referred to as the **permittivity of free space** and the magnetic field constant, μ_0, is referred to as the **permeability of free space**.

All physics students know this list as **Maxwell's equations**. When they were first published, they were called Heaviside's equations (or sometimes the Heaviside-Hertz equations since Heinrich Hertz discovered the same list simultaneously). Unfortunately, politics tend to play a role in how these things turn out and Heaviside was somewhat under-appreciated in his time, very much like Nikola Tesla. Many scientists felt that, since Maxwell was the first to try to unify electricity and magnetism, he should be given credit and so then they were called the Heaviside-Maxwell equations. In 1940, Albert Einstein published an article called *The Fundamentals of Theoretical Physics* where he referred them simply as Maxwell's equations and, from that point on, Heaviside's name has been lost in history.

5.5 Electromagnetic Waves

Maxwell's contributions to science are not limited to his edited Ampére's law. In the paper *A Dynamical Theory of the Electromagnetic Field*, he presented a derivation using his equations that showed electromagnetic waves could exist and traveled at the speed of light. Already knowing by experiment that light was affected by electric and magnetic fields, he concluded that light *was* an electromagnetic wave!

Maxwell's derivation was a bit involved because his list had so many equations. We'll use Heaviside's list (what we now call Maxwell's equations) to derive it in a much more succinct way just as Heinrich Hertz did. We know that light propagates through empty space where there is no charge or current. Therefore, we can write Eq. Set 5.4.9 as

$$\vec{\nabla} \bullet \vec{E} = 0 \tag{5.5.1a}$$

$$\vec{\nabla} \bullet \vec{B} = 0 \tag{5.5.1b}$$

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{5.5.1c}$$

$$\vec{\nabla} \times \vec{B} = \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \tag{5.5.1d}$$

because $\rho = 0$ and $\vec{J} = 0$ in empty space (remember, Maxwell's equations in del form apply to arbitrary points not whole spaces). Now, lets focus our attention on Eqs. 5.5.1c and 5.5.1d. If we take the curl of each of these, we get

$$\left\{ \begin{array}{l} \vec{\nabla} \times \left(\vec{\nabla} \times \vec{E} \right) = \vec{\nabla} \times \left(-\dfrac{\partial \vec{B}}{\partial t} \right) \\ \vec{\nabla} \times \left(\vec{\nabla} \times \vec{B} \right) = \vec{\nabla} \times \left(\mu_0 \epsilon_0 \dfrac{\partial \vec{E}}{\partial t} \right) \end{array} \right\}.$$

Since spatial derivative operators are commutative with time derivative operators, we get

$$\left\{ \begin{array}{l} \vec{\nabla} \times \left(\vec{\nabla} \times \vec{E} \right) = -\dfrac{\partial}{\partial t} \left(\vec{\nabla} \times \vec{B} \right) \\ \vec{\nabla} \times \left(\vec{\nabla} \times \vec{B} \right) = \mu_0 \epsilon_0 \dfrac{\partial}{\partial t} \left(\vec{\nabla} \times \vec{E} \right) \end{array} \right\}.$$

Using Eq. 3.2.8, we can substitute on the left side of the equations, which results in

$$\left\{ \begin{array}{l} \vec{\nabla}\left(\vec{\nabla} \bullet \vec{E}\right) - \vec{\nabla}^2\vec{E} = -\dfrac{\partial}{\partial t}\left(\vec{\nabla} \times \vec{B}\right) \\ \vec{\nabla}\left(\vec{\nabla} \bullet \vec{B}\right) - \vec{\nabla}^2\vec{B} = \mu_0\epsilon_0\dfrac{\partial}{\partial t}\left(\vec{\nabla} \times \vec{E}\right) \end{array} \right\}.$$

Inside each of the four sets of parentheses, we can substitute from Eq. Set 5.5.1 to arrive at

$$\left\{ \begin{array}{l} 0 - \vec{\nabla}^2\vec{E} = -\dfrac{\partial}{\partial t}\left(\mu_0\epsilon_0\dfrac{\partial\vec{E}}{\partial t}\right) \\ 0 - \vec{\nabla}^2\vec{B} = \mu_0\epsilon_0\dfrac{\partial}{\partial t}\left(-\dfrac{\partial\vec{B}}{\partial t}\right) \end{array} \right\}$$

$$\left\{ \begin{array}{rcl} \vec{\nabla}^2\vec{E} & = & \mu_0\epsilon_0\dfrac{\partial^2\vec{E}}{\partial t^2} \\ \vec{\nabla}^2\vec{B} & = & \mu_0\epsilon_0\dfrac{\partial^2\vec{B}}{\partial t^2} \end{array} \right\}. \tag{5.5.2}$$

These two equations match the form of the standard mechanical wave equation given by

$$\frac{d^2y}{dx^2} = \frac{1}{v^2}\frac{d^2y}{dt^2} \tag{5.5.3}$$

where we have a second derivative with respect to space proportional to a second derivative with respect to time. The proportionality constant is an inverse square of the wave-speed. This would suggest we can find the speed of an electromagnetic wave by stating

$$\frac{1}{c^2} = \mu_0\epsilon_0 \quad \Rightarrow c = \frac{1}{\sqrt{\mu_0\epsilon_0}}. \tag{5.5.4}$$

where c has the value of 299,792,458 m/s when you plug in the values of μ_0 and ϵ_0. This is the speed of light! This is also sometimes specified to be "in a vacuum" or "in free space" because experimentally (or practically) we

measure the speed of light to be different in different materials. In reality, however, light is never really *in* a material. Some of the light simply makes the atoms in the material emit their own light. These new light waves interfere with the original wave and the overall composite wave is all that we get to observe. It's complicated, but the point is the light only *appears* to travel at a speed slower than c.

According to Eq. 5.5.3, waves are a physical disturbance in a some medium represented by $y(x,t)$ where x represents the position of an arbitrary point in the medium. Based on Eq. 5.5.2, we can conclude that light is a disturbance in the electric and magnetic fields that exist throughout the universe. We have replaced the disturbance y measured in meters with a disturbance $\vec{E}$ or $\vec{B}$ measured in their respective units. In other words, $\vec{E}$ or $\vec{B}$ do not represent the fields already present at each point. They represent the amount by which those fields have been altered. The fields already present prior to the passage of the wave represent the equilibrium field strength, which we define as zero for waves.

This brings us to a question: How does one generate an electromagnetic wave? Well, it seems logical that, even though EM waves travel through empty space, they must have started somewhere that wasn't empty. They don't just appear out of nowhere (at least in the classical model). If we take another look at Maxwell's equations given by Eq. Set 5.4.9, then we see the source of our EM waves. Ampére's law says a changing electric field generates a magnetic field and Faraday's law says a changing magnetic field generates an electric field. Gauss's law says charges generate electric fields, so we can generate a changing electric field by moving some charges. However, this will only generate a static magnetic field and we need it to be changing. Under this logic, not only do the charges have to move, they have to change their motion so the magnetic field they generate also changes. A change of motion is given by an acceleration, so the logical conclusion is that accelerating charges generate electromagnetic waves!

Example 5.5.1

Just as with Eq. 5.5.3, there is a multitude of possible solutions to Eq. 5.5.2 involving the superposition of functions (in this case vector functions). The simplest of these solutions (worth examining) is for the **linearly-polarized**

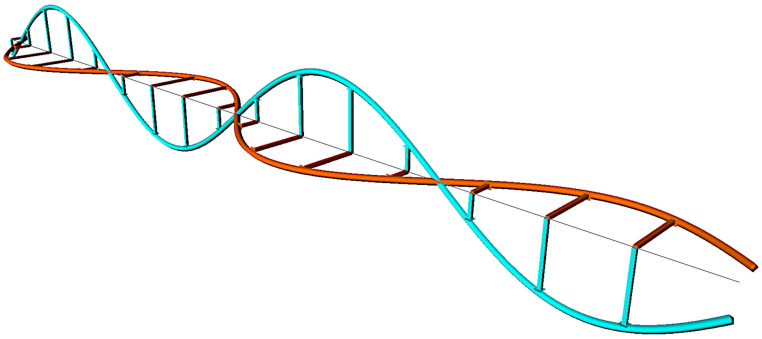


Figure 5.12: This is an example of an electromagnetic wave. Specifically, this type is called a plane linearly-polarized wave in which all vectors are oriented at 90°. The direction of propagation is downward to the right along the thin center line in the image.

plane wave shown in Figure 5.12. The solutions take the form

$$\left\{\begin{array}{l} \vec{E}(\vec{r},t) = \vec{E}_0 \cos\Big(\omega t - \vec{k} \bullet \vec{r} + \varphi_0\Big) \\ \vec{B}(\vec{r},t) = \vec{B}_0 \cos\Big(\omega t - \vec{k} \bullet \vec{r} + \varphi_0\Big) \end{array}\right\}, \tag{5.5.5}$$

where $\vec{r}$ is the position vector of the point in space, t is time, $\omega = 2\pi f$ is the angular frequency of the wave (in radians per second), $\vec{k} = (2\pi/\lambda)\,\hat{k}$ is the angular wave vector (in radians per meter) in the direction of propagation, and φ_0 is the phase angle (in radians). The vector quantities $\vec{E}_0$ and $\vec{B}_0$ are the corresponding amplitudes (maximum field disturbances) for each type of field.

Let's apply Eqs. 5.5.1a and 5.5.1c to these wave solutions. Assuming the direction of propagation is along the z-axis in Cartesian coordinates, we can say $\vec{k} \bullet \vec{r} = kz$ because of Eq. 1.1.1. Starting with Eq. 5.5.1a, we get

$$\vec{\nabla} \bullet \vec{E} = 0$$

$$\vec{\nabla} \bullet \Big[\vec{E}_0 \cos(\omega t - kz + \varphi_0)\Big] = 0$$

$$0 + 0 + \frac{\partial}{\partial z}\left[E_{0z} \cos(\omega t - kz + \varphi_0)\right] = 0$$

$$E_{0z} k \sin(\omega t - kz + \varphi_0) = 0.$$

Since $k \neq 0$ and $\sin(\omega t - kz + \varphi_0)$ cannot be zero *everywhere*, we can conclude $E_{0z} = 0$. This means the electric field disturbance of a linearly-polarized plane light wave is always orthogonal to the direction of propagation. In vector algebra terms, $\vec{E}_0 \bullet \vec{k} = 0$.

For the sake of simplicity, let's say the direction of $\vec{E}_0$ is $\hat{y}$. We can do this based on what we just stated because $\hat{y} \bullet \hat{z} = 0$ and $\vec{k} = k\hat{z}$. Starting with Eq. 5.5.1c results in

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

$$\vec{\nabla} \times \left[\vec{E}_0 \cos(\omega t - kz + \varphi_0)\right] = -\frac{\partial}{\partial t}\left[\vec{B}_0 \cos(\omega t - kz + \varphi_0)\right]$$

$$-\frac{\partial}{\partial z}\left[E_0 \cos(\omega t - kz + \varphi_0)\right]\hat{x} - 0 + 0 = -\frac{\partial}{\partial t}\left[\vec{B}_0 \cos(\omega t - kz + \varphi_0)\right]$$

$$-E_0 k \sin(\omega t - kz + \varphi_0)\,\hat{x} = \vec{B}_0 \omega \sin(\omega t - kz + \varphi_0)$$

$$\vec{B}_0 = -E_0 \frac{k}{\omega}\hat{x}.$$

It's in the $-\hat{x}$ direction. Therefore, the direction of the magnetic field disturbance of a plane linearly-polarized light wave is always orthogonal to the direction of propagation *and* the direction of the electric field disturbance. Furthermore,

$$B_0 = E_0 \frac{k}{\omega} = E_0 \frac{2\pi/\lambda}{2\pi f} = E_0 \frac{1}{\lambda f} = \frac{E_0}{c}$$

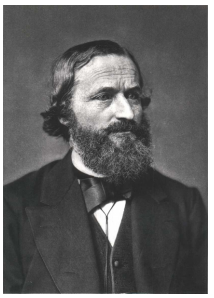
or sometimes written $\boxed{E_0 = c\,B_0}$. Not only are their directions related, so are their magnitudes.

In general, both field disturbances are orthogonal to the direction of propagation, but not necessarily to each other. We represent this fact by something called the **Poynting Vector** given by

$$\vec{S} = \frac{1}{\mu_0}\vec{E} \times \vec{B} \tag{5.5.6}$$

which is defined as the energy flux vector (in watts per square meter) of the EM wave. In other words, it's the rate of energy transfer per unit area in the direction of propagation.


Georg Ohm


Gustav Kirchhoff


Siméon Poisson

Figure 5.13: These people were important in the development of the electric potential.

5.6 Potential Functions

In Section 5.4, we introduced two quantities, ϕ and $\vec{A}$. Oliver Heaviside referred to these as a "physical inanity" (i.e. lacking physical substance). As it turns out, they are very closely tied to **energy**, a very physically significant quantity. However, to those like Heaviside in the mid-to-late 1800s, energy was a very new concept. Remember, we stated in Section 4.1, the principle of conservation of energy wasn't stated explicitly until 1845 by Hermann von Helmholtz. Energy can also seem a bit magical at times, so we can understand why, under these circumstance, Heaviside may have taken the stance that he did.

In purely mathematical terms, ϕ is called the **scalar potential** and $\vec{A}$ is called the **vector potential**. They are governed by a division of mathematics called Potential Theory. In the context of electrodynamics, ϕ is called the **electric potential** and $\vec{A}$ is called the **magnetic vector potential**. They are related to electric and magnetic fields through the del operator by

$$\vec{E} = -\vec{\nabla}\phi - \frac{\partial \vec{A}}{\partial t} \tag{5.6.1}$$

and

$$\vec{B} = \vec{\nabla} \times \vec{A}\,. \tag{5.6.2}$$

The first term in Eq. 5.6.1 matches what we know about scalar potentials for conservative fields (just as we saw with Eq. 4.2.3). As we can see, vector potentials are a bit trickier. The magnetic field is clearly the curl of $\vec{A}$ as

defined in Section 3.2. However, we can also see that a time-varying $\vec{A}$ can contribute to the overall electric field, a phenomenon that is easily described by Faraday's law (Eq. 5.4.9c).

Magnetostatics

If we assume for the moment that $\vec{A}$ is constant in time, then we have what we call the **magnetostatic approximation** (i.e. the study of static magnetic fields). This is an approximation we've already made in Section 5.3 without even realizing it. With this in mind, Eq. 5.6.1 becomes simply

$$\vec{E} = -\vec{\nabla}\phi \tag{5.6.3}$$

and we can say the electric field is a **conservative field** meaning it is path-independent. From this special case, we can form an argument for the physical significance of the electric potential. Evaluating Eq. 5.6.3 over a line integral from point a to point b, we get

$$\int_a^b \vec{E} \bullet d\vec{\ell} = -\int_a^b \vec{\nabla}\phi \bullet d\vec{\ell}.$$

The right side of this equation is just the fundamental theorem of vector calculus (Eq. 3.4.4), so

$$\int_a^b \vec{E} \bullet d\vec{\ell} = -\int_a^b d\phi$$

$$\int_a^b \vec{E} \bullet d\vec{\ell} = -\left[\phi|_b - \phi|_a\right]$$

$$\int_a^b \vec{E} \bullet d\vec{\ell} = \phi|_a - \phi|_b\,. \tag{5.6.4}$$

Therefore, the path integral of the electric field is just the difference in potential (or the **potential difference**) between the two endpoints a and b. Remember Faraday's law in integral form from 1831? The left side of Eq. 5.3.8 has a very similar integral form, which is no coincidence. A changing magnetic flux induced what Faraday called an **electromotive force** (or emf).

If we substitute Eq. 5.6.3 into Gauss's law (Eq. 5.4.9a), then we get

$$\vec{\nabla} \bullet \vec{E} = \frac{\rho}{\epsilon_0}$$

$$\vec{\nabla} \bullet \left(-\vec{\nabla}\phi\right) = \frac{\rho}{\epsilon_0}$$

$$\vec{\nabla}^2\phi = -\frac{\rho}{\epsilon_0}, \tag{5.6.5}$$

which is called **Poisson's equation** named for Siméon Denis Poisson. In free space where there is no charge, this takes the form

$$\vec{\nabla}^2\phi = 0, \tag{5.6.6}$$

which is called **Laplace's equation** named for Pierre-Simon Laplace (he did a lot for electrodynamics). Eq. 5.6.6 is applicable in quite a few unrelated fields (e.g. Thermodynamics), but is most noted in electrodynamics. The second space derivative operator on the left of Eqs. 5.6.5 and 5.6.6 is referred to as the **laplacian** (see Section 3.2)for reasons which should now be obvious.

In this magnetostatic case, the solution to Eq. 5.6.5 is given by an equation similar to Coulomb's law (Eq. 5.2.7):

$$d\phi = k_E \frac{dq}{r} = k_E \frac{\rho}{r}\, dV \tag{5.6.7}$$

or more specifically

$$d\phi = k_E \frac{\rho(t, \vec{r}_q)}{|\vec{r}_p - \vec{r}_q|}\, dV_q. \tag{5.6.8}$$

It may not have been obvious at the time, but a similar relation was found for $\vec{A}$ in Eq. 5.3.6. Taking note of Eq. 5.6.2, we get

$$d\vec{A} = k_M \frac{\vec{J}}{r}\, dV \tag{5.6.9}$$

or more specifically

$$d\vec{A} = k_M \frac{\vec{J}(\vec{r}_I)}{|\vec{r}_p - \vec{r}_I|}\, dV_I. \tag{5.6.10}$$

These equations are assuming that both charge density, ρ, and current density, $\vec{J}$, go to zero at infinity as they should in the real universe. In approximations that violate this, we have to be a little more creative.

Gauge Invariance

In 1848, Gustav Kirchhoff showed the electric potential, ϕ, to be the same as the "electric pressure" in Georg Simon Ohm's law regarding electric circuits (published in 1827). We now refer to this quantity as **voltage**. This is a fact Heaviside was well aware of, but still opted for vector fields $\vec{E}$ and $\vec{B}$ because the value of zero always meant something physical. The same cannot be said when ϕ and $\vec{A}$ have a value of zero.

The potential functions can vary by particular factors and still leave the vector fields $\vec{E}$ and $\vec{B}$ unchanged. This is called **gauge invariance**. The act of choosing a gauge is called **gauge fixing** and it allows us to not only be speaking the same language, but also simplify equations a bit. The gauge invariance for electrodynamic potentials is given by

$$\phi \rightarrow \phi - \frac{\partial f}{\partial t} \tag{5.6.11a}$$

$$\vec{A} \rightarrow \vec{A} + \vec{\nabla} f \tag{5.6.11b}$$

where $f(t, \vec{r})$ is an arbitrary **gauge function**. We can substitute Eq. Set 5.6.11 into Eq. 5.6.1,

$$\begin{aligned} \vec{E} &= -\vec{\nabla}\left(\phi - \frac{\partial f}{\partial t}\right) - \frac{\partial}{\partial t}\left(\vec{A} + \vec{\nabla} f\right) \\ \vec{E} &= -\vec{\nabla}\phi + \vec{\nabla}\left(\frac{\partial f}{\partial t}\right) - \frac{\partial \vec{A}}{\partial t} - \frac{\partial}{\partial t}\left(\vec{\nabla} f\right). \end{aligned}$$

Since the del operator and the time derivative are commutative, the mixed terms cancel leaving us with just Eq. 5.6.1. We can make similar substitutions in Eq. 5.6.2 arriving at

$$\vec{B} = \vec{\nabla} \times \left(\vec{A} + \vec{\nabla} f\right) = \vec{\nabla} \times \vec{A} + \vec{\nabla} \times \vec{\nabla} f.$$

Since the curl of the gradient is always zero (Eq. 3.2.6), the second term disappears and we get just Eq. 5.6.2.

Gauges in physics are not usually defined by specifying a function f, but rather by specifying the divergence of $\vec{A}$. Eqs. 5.6.1 and 5.6.2 say nothing about how $\vec{A}$ diverges and so it is an arbitrary quantity. There are a couple very popular gauges: the **Coulomb gauge**, given by

$$\vec{\nabla} \bullet \vec{A} = 0, \tag{5.6.12}$$

and the **Lorenz gauge** (not to be confused with Lorentz), given by

$$\vec{\nabla} \bullet \vec{A} = -\frac{1}{c^2}\frac{\partial \phi}{\partial t}. \tag{5.6.13}$$

These have particular uses when applying them to Maxwell's equations.

Maxwell's Equations with Potentials

We can write Maxwell's equations (Eq. Set 5.4.9) entirely in terms of potentials using Eqs. 5.6.1 and 5.6.2. The result is astonishing because two of them, Eqs. 5.4.9b and 5.4.9c, are automatically satisfied:

$$\vec{\nabla} \bullet \vec{B} = \vec{\nabla} \bullet \left(\vec{\nabla} \times \vec{A}\right) = 0$$

because the divergence of a curl is always zero (Eq. 3.2.7) and

$$\begin{aligned}
\vec{\nabla} \times \vec{E} &= \vec{\nabla} \times \left(-\vec{\nabla}\phi - \frac{\partial \vec{A}}{\partial t}\right) \\
\vec{\nabla} \times \vec{E} &= -\vec{\nabla} \times \left(\vec{\nabla}\phi\right) - \vec{\nabla} \times \left(\frac{\partial \vec{A}}{\partial t}\right) \\
\vec{\nabla} \times \vec{E} &= -\frac{\partial}{\partial t}\left(\vec{\nabla} \times \vec{A}\right) = -\frac{\partial \vec{B}}{\partial t}
\end{aligned}$$

because the curl of a gradient is always zero (Eq. 3.2.6) and del is commutative with a time derivative. Because they're automatically satisfied, we don't even have to list them!

Eqs. 5.4.9a and 5.4.9d are a bit more involved. Eq. 5.4.9a becomes

$$\begin{aligned}
\vec{\nabla} \bullet \vec{E} &= \frac{\rho}{\epsilon_0} \\
\vec{\nabla} \bullet \left(-\vec{\nabla}\phi - \frac{\partial \vec{A}}{\partial t}\right) &= \frac{\rho}{\epsilon_0} \\
-\vec{\nabla} \bullet \left(\vec{\nabla}\phi\right) - \vec{\nabla} \bullet \left(\frac{\partial \vec{A}}{\partial t}\right) &= \frac{\rho}{\epsilon_0}
\end{aligned}$$

$$-\vec{\nabla}^2\phi - \frac{\partial}{\partial t}\left(\vec{\nabla} \bullet \vec{A}\right) = \frac{\rho}{\epsilon_0}. \tag{5.6.14}$$

This is where the gauge fixing comes into play. Under the Coulomb gauge (Eq. 5.6.12), we get

$$\vec{\nabla}^2\phi = -\frac{\rho}{\epsilon_0},$$

which is just Poisson's equation (Eq. 5.6.5) just like with magnetostatics. The coulomb gauge does make it particularly easy to find the electric potential, but $\vec{A}$ is still rather challenging. In this more general case, ϕ is not enough to determine $\vec{E}$ (see Eq. 5.6.1), so $\vec{A}$ must be found. Furthermore, changes in ϕ over time propagate through space instantaneously, which is still physically legal because ϕ is not a physically measurable quantity. At this moment, you might be yelling at this book saying "I've measured potential before!". The truth is you've never measured potential. You haven't even measured $\vec{E}$. What you do measure is the effect $\vec{E}$ has on physical objects and you interpret this as a ϕ or an $\vec{E}$. Since $\vec{E}$ is also dependent on $\vec{A}$ and changes in $\vec{A}$ propagate at the speed of light, we're not violating any physical laws.

Under the Lorenz gauge (Eq. 5.6.13), things are a bit simpler overall. Eq. 5.6.14 becomes

$$-\vec{\nabla}^2\phi - \frac{\partial}{\partial t}\left(-\frac{1}{c^2}\frac{\partial\phi}{\partial t}\right) = \frac{\rho}{\epsilon_0}$$

$$\vec{\nabla}^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = -\frac{\rho}{\epsilon_0}. \tag{5.6.15}$$

This might seem a bit more complicated, but now changes in ϕ over time only propagate at the speed of light, so it makes more sense. The Lorenz gauge also simplifies Eq. 5.4.9d to

$$\begin{aligned}
\vec{\nabla} \times \vec{B} &= \mu_0\vec{J} + \mu_0\epsilon_0\frac{\partial\vec{E}}{\partial t} \\
\vec{\nabla} \times \left(\vec{\nabla} \times \vec{A}\right) &= \mu_0\vec{J} + \mu_0\epsilon_0\frac{\partial}{\partial t}\left(-\vec{\nabla}\phi - \frac{\partial\vec{A}}{\partial t}\right).
\end{aligned}$$

By Eq. 3.2.8, we get

$$\begin{aligned}\vec{\nabla}\left(\vec{\nabla}\bullet\vec{A}\right)-\vec{\nabla}^2\vec{A} &= \mu_0\vec{J}+\mu_0\epsilon_0\frac{\partial}{\partial t}\left(-\vec{\nabla}\phi-\frac{\partial\vec{A}}{\partial t}\right)\\ \vec{\nabla}\left(\vec{\nabla}\bullet\vec{A}\right)-\vec{\nabla}^2\vec{A} &= \mu_0\vec{J}-\mu_0\epsilon_0\frac{\partial}{\partial t}\left(\vec{\nabla}\phi\right)-\mu_0\epsilon_0\frac{\partial^2\vec{A}}{\partial t^2}.\end{aligned}$$

Since the divergence of $\vec{A}$ is once again given by our gauge, we have

$$\vec{\nabla}\left(-\frac{1}{c^2}\frac{\partial\phi}{\partial t}\right)-\vec{\nabla}^2\vec{A}=\mu_0\vec{J}-\mu_0\epsilon_0\frac{\partial}{\partial t}\left(\vec{\nabla}\phi\right)-\mu_0\epsilon_0\frac{\partial^2\vec{A}}{\partial t^2}.$$

We also know del is commutative with time and the speed of light, c, is given by Eq. 5.5.4. Therefore, we have

$$-\frac{1}{c^2}\frac{\partial}{\partial t}\left(\vec{\nabla}\phi\right)-\vec{\nabla}^2\vec{A}=\mu_0\vec{J}-\frac{1}{c^2}\frac{\partial}{\partial t}\left(\vec{\nabla}\phi\right)-\frac{1}{c^2}\frac{\partial^2\vec{A}}{\partial t^2}.$$

The first term on the left cancels with the second term on the right.

$$-\vec{\nabla}^2\vec{A}=\mu_0\vec{J}-\frac{1}{c^2}\frac{\partial^2\vec{A}}{\partial t^2}$$

$$\vec{\nabla}^2\vec{A}-\frac{1}{c^2}\frac{\partial^2\vec{A}}{\partial t^2}=-\mu_0\vec{J}. \tag{5.6.16}$$

Not only do Eqs. 5.6.15 and 5.6.16 retain the beautiful symmetry of Maxwell's equations, but they also very quickly show wave equations in free space for light. The only downside to writing Maxwell's equations this way is that we're dealing with second-order differential equations rather than first order ones. Having to keep track of a gauge may be something people like Oliver Heaviside didn't want to do, but we'll see in a later chapter that we can show the electric and magnetic vector potentials to be more physical than the electric and magnetic fields.

5.7 Blurring Lines

Through this entire chapter we've been discussing E-fields and B-fields as if they're entirely separate entities. However, the very name of Section 5.4

should be an indication they are not. In that section, we listed 5 equations that describe entirely the discipline of electrodynamics. They were Eq. Set 5.4.9 and Eq. 5.4.10. Eq. 5.4.10 is of particular interest to us in this section. We'll rewrite it as

$$\vec{F} = q\left(\vec{E} + \vec{v} \times \vec{B}\right) \tag{5.7.1}$$

by factoring out the charge q.

Back in Section 5.2, we explained that $\vec{E}$ and $\vec{B}$ were just mathematical middle-men used to simplify our model of how charges interact. These two fields are purely mathematical (i.e. not physical) quantities. Even in Section 5.6, we explained that you've never actually measured them before. In case what I'm trying to say still isn't entirely clear, I'll say it as succinct as I can: "Fields are *not* real!" You might ask "So what is real then?" The answer is "the response of the charges."

Charges respond to each other by accelerating. We know from Newton's second law (Eq. 4.2.6) that acceleration is directly proportional to net force, which brings us to Eq. 5.7.1. The parenthetical quantity $\vec{E} + \vec{v} \times \vec{B}$ can be referred to as the **electromagnetic field**. Richard Feynman once said "One part of the force between moving charges we call the magnetic force. It is really one aspect of an electrical effect." The truth is that $\vec{E}$ and $\vec{B}$ really just represent two aspects of the same idea: effects on charges. Furthermore, which is which really depends on your point of view. The $\vec{v} \times \vec{B}$ term is velocity-dependent and we know from classical mechanics that velocity is relative to the observer (a concept around since Galileo). If the point of view says $\vec{v} = 0$, then the $\vec{E}$ term is going to have to make up for that lost effect. All observers should measure the same acceleration and, therefore the same force given by Eq. 5.7.1 (at least for $v \ll c$).

Electrodynamics is simply a model for how charges interact with one another. The two fields we use are a mathematical tool to make the model more practical. In the end, it is just a model and only a classical one at that.

Chapter 6

Tensor Analysis

6.1 What is a Tensor?

The simplest explanation of a **tensor** is that it's a way of combining similar quantities to simplify a set of mathematics, but it's a bit more than that. The word "tensor" refers to a more specific type of combined quantity. There are quantities called **pseudotensors** that look like tensors and behave *almost* like tensors, but are not quite tensors. Before we can properly define a tensor, we need to get a solid grip on the notation we use to represent them.

6.2 Index Notation

The most common way to represent a tensor is to use an index and operate by components. How many indices the tensor has tells us the tensor's **rank**. We have already used tensors of lower rank without even realizing it. For example,

- a tensor of rank-0, T, is a scalar.
  (i.e. no values are required to determine the component.)
- a tensor of rank-1, T_i, is a vector.
  (i.e. only one value is required to determine the component.)

Tensors of higher rank are called dyads (rank-2 T_{ij}), triads (rank-3 T_{ijk}), quadads (rank-4 T_{ijkl}), etc. However, these names are seldom used. The number of values each index can take tells us the tensor's **dimension**. For

example, a vector of dimension-3 like T_i will have the components T_1, T_2, and T_3. Likewise, a vector of dimension-4 will have 4 components. A rank-2 tensor of dimension-3 will have $3^2 = 9$ components and a rank-2 tensor of dimension-4 will have $4^2 = 16$ components. We can state this in general by saying a rank-n tensor of dimension-m has m^n components. Sometimes, we distinguish between dimensions by using latin letters for dimension-3 and greek letters for dimension-4. This is a convention I have adopted for this book.

Each tensor component is given in terms of a set of coordinates. These coordinates come in two forms: **covariant** and **contravariant**. In abstract mathematics, these two types of coordinates are very distinct. However, in practical situations such as physics, we use orthogonal (often orthonormal) coordinates. In the special case where all coordinates are orthogonal, the difference between covariance and contravariance blurs significantly. In fact, they're identical if we further simplify to Cartesian 3-space.

A covariant coordinate, x_i, is shown by using a lower index and a contravariant coordinate, x^i, is shown by using an upper index. As I'm sure you've noticed, a contravariant coordinate index can be easily confused with an exponent. To compensate, we try to avoid using exponents in index notation (e.g. x^2 would be written as xx instead). Tensors written in terms of these coordinates have a similar notation. In order for a tensor to be covariant, all its indices must be lower. Likewise, for a tensor to be contravariant, all its indices must be upper. Otherwise, the tensor is considered **mixed**. For example,

- T_i is a covariant vector.
- T^i is a contravariant vector.
- T_{ij} is a covariant rank-2 tensor.
- T^{ij} is a contravariant rank-2 tensor.
- T^i_j is a mixed rank-2 tensor.

This pattern continues for higher rank tensors.

Another convention used with this notation is called the **Einstein summation convention**, which is applied a great deal in Einstein's General Theory of Relativity. Operations between tensors often involve a summation

and writing the summation sign can get old fast, so we have a way of implying the summation instead. For example, let's take the 3-space dot product given by Eq. 2.2.2 as

$$\vec{A} \bullet \vec{B} = \sum_{i=1}^{3} A_i B_i = A_1 B_1 + A_2 B_2 + A_3 B_3.$$

Under the notational standards given in this section, however, one of these vectors should be covariant and the other contravariant. Therefore, the dot product is really

$$\vec{A} \bullet \vec{B} = \sum_{i=1}^{3} A^i B_i = A^1 B_1 + A^2 B_2 + A^3 B_3.$$

The Einstein summation convention states if an index is repeated, upper on one tensor and lower on another, then the summation is implied and we need not write the summation symbol. We can now write the dot product simply as

$$\vec{A} \bullet \vec{B} = A^i B_i = A^1 B_1 + A^2 B_2 + A^3 B_3 \tag{6.2.1}$$

where the index i is repeated (i.e. summed over) and the vectors are dimension-3 implied by the use of latin letters.

Example 6.2.1

When we're first introduced to the moment of inertia, it's defined as a measure of an object's ability to resist changes in rotational motion. We're also given little formulae which all depend on mass and, more importantly, the mass distribution. However, in general, moment of inertia also depends on the orientation of the rotational axis and the best way to represent such ambiguity is with a tensor.

In order to find the form of this tensor in index notation, we'll start with the origin of the moment of inertia: spin angular momentum. Spin angular momentum is given by

$$\vec{L}_{\text{spin}} = \sum \vec{r} \times \vec{p} = \sum m\vec{r} \times \vec{v}$$

where $\vec{r}$ and $\vec{v}$ are the position and velocity, respectively, of a point mass m relative to the center of mass of the body. If the body has enough m's closely

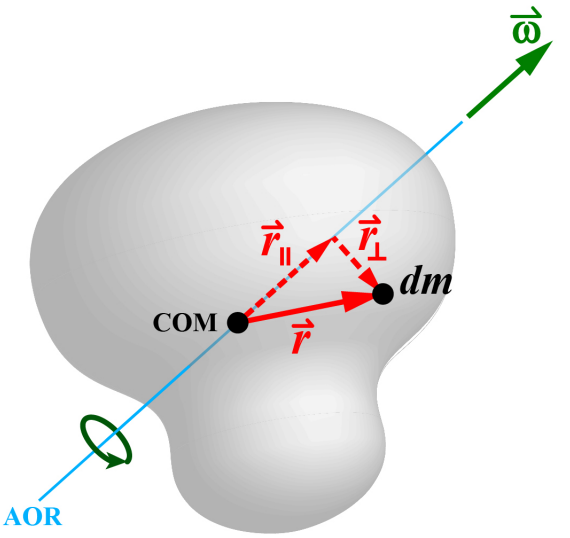


Figure 6.1: This is an arbitrary rigid body. Its center of mass (COM), axis of rotation (AOR), and mass element (dm) have been labeled. The position of dm relative to the COM is given by $\vec{r}$.

packed, then we can treat the body as continuous. Under those conditions, spin angular momentum is

$$\vec{L}_{\text{spin}} = \int \vec{r} \times \vec{v}\, dm$$

where dm is the mass element of the body.

Typically, when discussing moments of inertia, we're dealing with rigid bodies. A rigid body is one in which each r (i.e. the magnitude of $\vec{r}$) does not change in time. As shown in Figure 6.1, each mass element travels in a circle of radius $r_\perp$ around the axis of rotation and an angular velocity $\vec{\omega}$ is common to all mass elements. Therefore, the velocity of the mass element is given by

$$\vec{v} = \vec{\omega} \times \vec{r_\perp} = \vec{\omega} \times \left(\vec{r} - \vec{r}_{||}\right) = \vec{\omega} \times \vec{r} - \vec{\omega} \times \vec{r}_{||}.$$

Since both $\vec{\omega}$ and $\vec{r}_{||}$ are parallel to the rotational axis, their cross product is zero according to Eq. 2.2.3 and the velocity of each mass element becomes

$$\vec{v} = \vec{\omega} \times \vec{r}.$$

Substituting this back into the spin angular momentum, we get

$$\vec{L}_{\text{spin}} = \int \vec{r} \times (\vec{\omega} \times \vec{r}) \; dm.$$

What we have here is a triple product which obeys the identity given by Eq. 2.2.12. Now the spin angular momentum can be written as

$$\vec{L}_{\text{spin}} = \int \left[\vec{\omega}\,(\vec{r} \bullet \vec{r}) - \vec{r}\,(\vec{r} \bullet \vec{\omega})\right] \; dm.$$

Dot products in index notation are given by Eq. 6.2.1, so we can write

$$L_i = \int \left[\omega_i r^k r_k - r_i r^j \omega_j\right] \; dm,$$

which is the i^{th} component of spin angular momentum. The index i is referred to as a **free index** where as j and k are each a **summation index**. All free indices on the left side of a tensor equation must match those on the right side in symbol and location.

We cannot simply pull out the ω because each one is indexed differently. The rank of each term must be maintained, so we need to use a special rank-2 mixed tensor given by

$$\delta^i_j = \begin{cases} 1, & \text{when } i = j \\ 0, & \text{when } i \neq j \end{cases} \tag{6.2.2}$$

which is called the **Kronecker delta**. With this tensor, we can say $\omega_i = \delta^j_i \omega_j$ and spin angular momentum becomes

$$\begin{aligned} L_i &= \int \left[\delta^j_i \omega_j r^k r_k - r_i r^j \omega_j\right] \; dm \\ L_i &= \left(\int \left[\delta^j_i r^k r_k - r_i r^j\right] dm\right) \omega_j. \end{aligned}$$

The parenthetical quantity can now be defined as

$$\boxed{I^j_i = \int \left[\delta^j_i r^k r_k - r_i r^j\right] dm}, \tag{6.2.3}$$

which is the moment of inertia tensor. This leaves us with a spin angular momentum of $L_i = I^j_i \omega_j$. The moment of inertia tensor is a rank-2 dimension-3 tensor. If the axis of rotation is a principle axis (i.e. an axis of symmetry) of the rigid body, then all components where $i \neq j$ will be zero.

6.3 Matrix Notation

Even though some generality is lost, it's sometimes a good visual to represent tensors using matrices since the operations are very similar. A scalar would be a single component matrix (e.g. $T = [2.73]$ K). A vector would be represented as a row or column matrix depending on the desired operation. For example,

$$\vec{v} = \begin{bmatrix} 2 & 3 & 5 \end{bmatrix} \frac{\text{m}}{\text{s}} \quad \text{or} \quad \vec{v} = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix} \frac{\text{m}}{\text{s}}$$

are dimension-3 velocity vectors.

Example 6.3.1

This matrix vector notation carries over into operations like the dot product given in Eq. 6.2.1. A common application of the dot product is work (as seen in Example 2.2.1) defined by

$$W = \int \vec{F} \bullet d\vec{s} = \int \vec{F} \bullet \vec{v}\, dt.$$

In index notation, this would be written as

$$W = \int F^i v_i \, dt.$$

We can also write the vectors $\vec{F}$ and $\vec{v}$ as matrices. In matrix notation, work becomes

$$W = \int \begin{bmatrix} F^1 & F^2 & F^3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} dt,$$

which by matrix operations would have exactly the same result as the standard dot product.

Don't be fooled by anyone claiming covariant vectors are always column matrices (and contravariant vectors are always row matrices). The dot product given in Eq. 6.2.1 is valid if it's written $A^i B_i$ or $B_i A^i$ and should still result in a scalar. In other words, the row matrix must always be written

first (regardless of "variance") because of the way matrices operate on one another. Similar issues arise elsewhere which can make matrix notation a bit cumbersome at times.

A rank-2 tensor is represented by a square matrix with a number of rows (as well as columns) equal to the dimension of the tensor. For example,

$$\sigma_{ij} \longrightarrow \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 8 \\ 0 & 2 & 0 \\ 8 & 0 & 3 \end{bmatrix} \frac{\mathrm{N}}{\mathrm{m}^2}$$

is a rank-2 dimension-3 covariant tensor. Specifically, this is an example of the Cauchy stress tensor where the diagonal components represent pressure and the off-diagonal components represent shear stress. This tensor is always symmetric across the diagonal in matrix notation (i.e. $\sigma_{ij} = \sigma_{ji}$). This particular example also represents the origin of the word "tensor" (tension). The long arrow in the above equation is used because an arbitrary component, σ_{ij}, cannot be *equal* to an entire tensor. It simply indicates a change in notation.

The Cauchy stress tensor is extended in General Relativity to dimension-4. This generalization is called the stress-energy tensor and is in the form

$$T_{\alpha\beta} \longrightarrow \begin{bmatrix} T_{00} & T_{01} & T_{02} & T_{03} \\ T_{10} & T_{11} & T_{12} & T_{13} \\ T_{20} & T_{21} & T_{22} & T_{23} \\ T_{30} & T_{31} & T_{32} & T_{33} \end{bmatrix}.$$

This tensor is symmetric and has discernible pieces. The lower right 3×3 is the Cauchy stress tensor, T_{00} is the energy density, $[T_{01}, T_{02}, T_{03}]$ is the energy flux vector, and $[T_{10}, T_{20}, T_{30}]$ is the momentum density vector (which, by symmetry, is the same as the energy flux vector). We'll get into the details later in the book.

Another example is the Kronecker Delta defined by Eq. 6.2.2 and given in matrix notation as

$$\delta^i_j \longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

This is simply the dimension-3 identity matrix. Its use is important because it is used to maintain rank when factoring an expression just as in Example 6.2.1.

Example 6.3.2

In Example 6.2.1, the final result was the equation $L_i = I_i^j \omega_j$, which in matrix notation is

$$\begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} = \begin{bmatrix} I_1^1 & I_1^2 & I_1^3 \\ I_2^1 & I_2^2 & I_2^2 \\ I_3^1 & I_3^2 & I_3^3 \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix}.$$

Operating using matrix multiplication results in

$$\begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} = \begin{bmatrix} I_1^1\omega_1 + I_1^2\omega_2 + I_1^3\omega_3 \\ I_2^1\omega_1 + I_2^2\omega_2 + I_2^3\omega_3 \\ I_3^1\omega_1 + I_3^2\omega_2 + I_3^3\omega_3 \end{bmatrix}$$

which has components in a form that match the original index notation. The index j is the summation index and each of these components is a summation over those indices.

If we wanted to isolate the moment of inertia tensor in matrix form, then we would need to decide on a coordinate system. Let's keep things simple and choose Cartesian. Based on Eq. 6.2.3, the moment of inertia is

$$I_i^j \longrightarrow \int \begin{bmatrix} \delta_1^1 x^k x_k - x_1 x^1 & \delta_1^2 x^k x_k - x_1 x^2 & \delta_1^3 x^k x_k - x_1 x^3 \\ \delta_2^1 x^k x_k - x_2 x^1 & \delta_2^2 x^k x_k - x_2 x^2 & \delta_2^3 x^k x_k - x_2 x^3 \\ \delta_3^1 x^k x_k - x_3 x^1 & \delta_3^2 x^k x_k - x_3 x^2 & \delta_3^3 x^k x_k - x_3 x^3 \end{bmatrix} dm.$$

Since only the diagonal components are non-zero in the Kronecker Delta, we have

$$I_i^j \longrightarrow \int \begin{bmatrix} x^k x_k - x_1 x^1 & -x_1 x^2 & -x_1 x^3 \\ -x_2 x^1 & x^k x_k - x_2 x^2 & -x_2 x^3 \\ -x_3 x^1 & -x_3 x^2 & x^k x_k - x_3 x^3 \end{bmatrix} dm.$$

Performing the summation over the index k results in

$$I_i^j \longrightarrow \int \begin{bmatrix} x_2 x^2 + x_3 x^3 & -x_1 x^2 & -x_1 x^3 \\ -x_2 x^1 & x_1 x^1 + x_3 x^3 & -x_2 x^3 \\ -x_3 x^1 & -x_3 x^2 & x_1 x^1 + x_2 x^2 \end{bmatrix} dm.$$

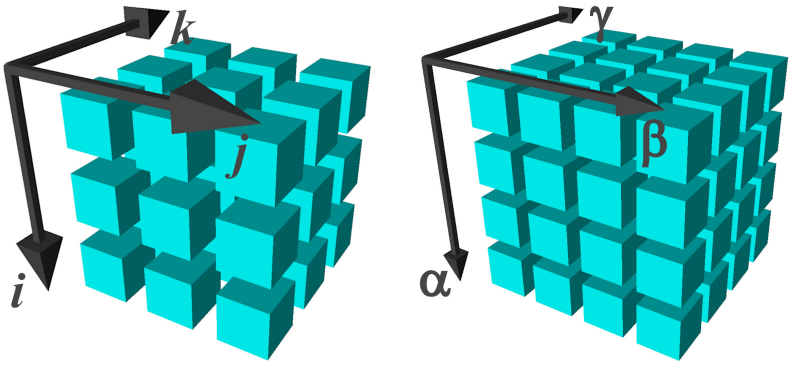
Now that we've performed all the operations associated with the indices, we can drop that notation entirely arriving at

$$\boxed{I_i^j \longrightarrow \int \begin{bmatrix} yy+zz & -xy & -xz \\ -yx & xx+zz & -yz \\ -zx & -zy & xx+yy \end{bmatrix} dm}$$

where $x_1 = x^1 \equiv x$, $x_2 = x^2 \equiv y$, and $x_3 = x^3 \equiv z$. Since we're working in Cartesian space, the covariant and contravariant coordinates are the same.

You might think the matrix notation ends with rank-2 tensors. However, while first learning about number arrays in high school computer programming class, I designed a visual representation for higher rank tensors akin to matrices. Let's consider the pattern developing here. A scalar (rank-0 tensor) has a single component, a vector (rank-1 tensor) has a length of components, and a rank-2 tensor has a length and width of components. It stands to reason that a rank-3 tensor should have a length, width, and depth of components like that given in Figure 6.2.

Rank-4 tensors, like those found all over General Relativity, might seem impossible under this pattern until you consider the subtle aspects. A rank-1 tensor is a collection of rank-0 tensors, a rank-2 is a collection of rank-1's, and a rank-3 is a collection of rank-2's. Therefore, I would argue that a rank-4 is simply a collection of rank-3's like that given in Figure 6.3. Unfortunately, we're beginning to see the problem with matrix notation. How does something like a rank-4 tensor operate?! It is usually best to yield to index notation and treat matrix notation as simply a way to visualize the quantity.



Rank-3 Dimension-3 Rank-3 Dimension-4

Figure 6.2: These are both rank-3 tensors in matrix notation. The tensor on the left is dimension-3 (T_{ijk}) and the tensor on the right is dimension-4 ($T_{\alpha\beta\gamma}$).

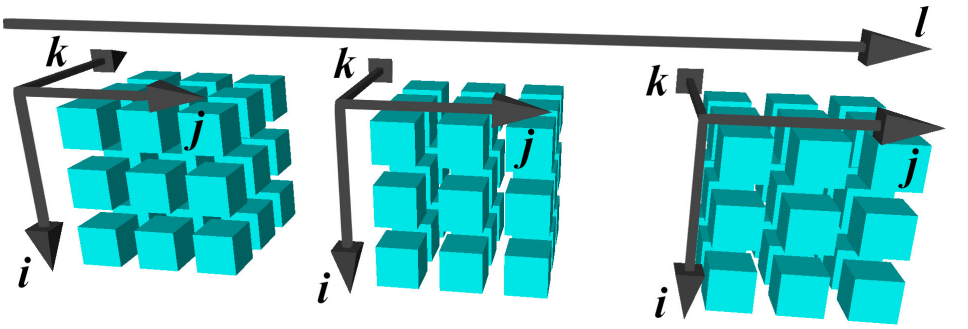


Figure 6.3: This is a rank-4 dimension-3 tensor in matrix notation. In index notation, it would be represented by T_{ijkl} where the final index l is given by the large axis on the left (i.e. it tells you which rank-3 you're in).

6.4 Describing a Space

As seen in Section 6.3, tensors are a great deal like matrices. Matrices had been combining similar quantities in mathematics for centuries before tensors were around, so why the new terminology? The truth is tensors are much more than just matrices. Tensors incorporate directional information through the use of coordinate systems and, as we saw in Chapter 1, there are quite a few to choose from.

Line Element

The simplest, most straight-forward way to represent a coordinate system with tensors is to use a scalar quantity called a **line element**. This line element describes the infinitesimal distance between two consecutive points in a space and will look different depending on the coordinate system choice. For example, in Cartesian three-space, the line element is

$$ds^2 = dx^2 + dy^2 + dz^2 \tag{6.4.1}$$

and, in spherical three-space, it's

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta \, d\phi^2, \tag{6.4.2}$$

where the 2's are exponents. With a careful look at Eq. 3.4.3, we can see that

$$ds^2 = d\vec{\ell} \bullet d\vec{\ell} = d\ell_j \, d\ell^j,$$

where $d\vec{\ell}$, the path element, is written in whatever coordinate system you may need.

Metric Tensor

Formally, we treat the scale factors (to use terminology from Section 3.4) separate from the coordinates x^i, so we'd like to separate these scale factors in the definition of the line element as well. This requires defining a new quantity called the **metric tensor**, g_{ij}. Now, the line element can be written

$$\boxed{ds^2 = g_{ij} \, dx^i \, dx^j}, \tag{6.4.3}$$

where both i and j are summation indices and x^i is a contravariant coordinate. By this definition, the metric tensor contains all information about the shape of the space. If the coordinate system choice changes, then g_{ij} must also change. Even more importantly, if the inherent shape of the space is changed, then g_{ij} must also change. This last statement is a hint at a discipline of mathematics called differential geometry, the foundation of General Relativity.

Eq. 6.4.3 is general enough to apply to all coordinate systems, but we can still write the metric tensor's components in specific coordinate systems. We use the definition

$$g_{ij} = \vec{e}_i \bullet \vec{e}_j = \left(\vec{e}_i\right)^k \left(\vec{e}_j\right)_k , \tag{6.4.4}$$

where $\vec{e}_j$ is a **coordinate basis** vector and $\left(\vec{e}_j\right)_k$ is the k^{th} component of that vector. In Cartesian coordinates, we have

$$g_{ij} = \delta_{ij} \longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} ; \tag{6.4.5}$$

and, in spherical coordinates, we have

$$g_{ij} \longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{bmatrix} \tag{6.4.6}$$

where the 2's are exponents. In each case, we see the tensor is diagonal with components equal to the square of the scale factor (e.g. $g_{\theta\theta} = \vec{e}_\theta \bullet \vec{e}_\theta = h_\theta h_\theta$). However, this is only the case when the space is described by orthogonal basis vectors (i.e. $\vec{e}_i \bullet \vec{e}_j = 0$ when $i \neq j$). The metric tensor may not be diagonal in general, but it is always symmetric since the dot product is commutative.

Raising and Lowering Indices

Beyond simply describing the space, the metric tensor also allows us to raise and lower indices on other tensors (i.e. convert between contravariant and covariant forms). For example, we can lower indices by

- $T_i = g_{ij} T^j$.

- $T_i^j = g_{ik}T^{kj}$.
- $T_{ij} = g_{ik}T^{kl}g_{lj}$.

This pattern continues for higher rank tensors. Raising indices requires the **inverse metric tensor**, which can be found using standard matrix algebra. For example, it is $g^{ij} = g_{ij}$ in Cartesian coordinates and

$$g^{ij} \longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{r^2} & 0 \\ 0 & 0 & \frac{1}{r^2 \sin^2\theta} \end{bmatrix} \tag{6.4.7}$$

in spherical coordinates. Using it we can raise indices by

- $T^i = g^{ij}T_j$.
- $T^i_j = g^{ik}T_{kj}$.
- $T^{ij} = g^{ik}T_{kl}g^{lj}$.

This pattern also continues for higher rank tensors. An interesting result of all this is $g^i_j = g^{ik}g_{kj} = \delta^i_j$, which makes sense if you think in terms of inverse matrices. Raising and lowering indices is very useful when writing complex tensor equations.

Coordinate Basis vs. Orthonormal Basis

A drawback to this form of the metric tensor is that we're using a **coordinate basis**, $\vec{e}_i$, as opposed to an **orthonormal basis**, $\hat{e}_i$. That means the basis vectors are all orthogonal, but not necessarily unit vectors (i.e. they don't necessarily have a magnitude of one). For example, in cylindrical coordinates, $\vec{e}_\phi = r\hat{e}_\phi = r\hat{\phi}$ meaning $\vec{e}_\phi$ has a magnitude of r (or is larger the further you are from the origin). This is something we're forced into if we wish to discuss space in terms of coordinates. Unfortunately, most basic physics is done in some kind of orthonormal basis. We can project onto one using

$$T_{\hat{k}\hat{l}} = (\hat{e}_k)^i \, (\hat{e}_l)^j \, T_{ij} \tag{6.4.8}$$

where $(\hat{e}_k)^i$ is the i^{th} coordinate basis component of the k^{th} orthonormal basis vector (meaning you'll need to write out the orthonormal basis vectors in the

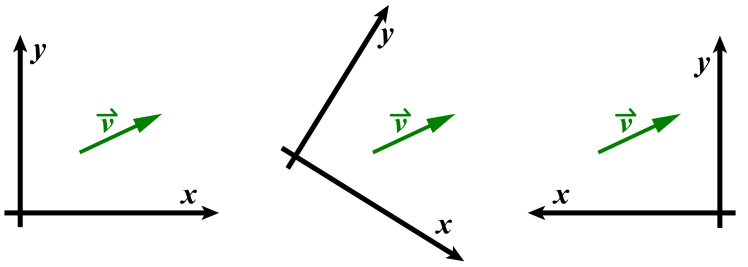


Figure 6.4: This diagram demonstrates how the fundamental nature of a vector remains unchanged when the coordinate system is rotated (center) or reflected (far right).

coordinate basis). Performing this process on the metric tensor always gives

$$g_{\hat{k}\hat{l}} = (\hat{e}_k)^i \, (\hat{e}_l)^j \, g_{ij} \longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{6.4.9}$$

which is just the metric tensor for Cartesian space. Sometimes, $g_{\hat{k}\hat{l}}$ is written as η_{kl}, but I find that much less descriptive and there are already enough symbols to worry about. This projection is usually a final step in any work, but must eventually be done to make real sense of your results especially if those results will be used in another physics discipline.

6.5 Really... What's a Tensor?!

At the beginning of this chapter, we mentioned a tensor was a special kind of quantity grouping. A common definition for the word "tensor" is a quantity which remains unchanged when transformed from one set of coordinates to *any* other set. Just to be clear, we don't mean completely unchanged because the only quantity that does that is a scalar. What we mean is that the physical nature of the tensor is unchanged.

A common example given when discussing tensors is the velocity vector. The components of velocity will change when a coordinate system is rotated, but the *motion* of the object is not changed by the transformation as shown in Figure 6.4. The velocity will point the same direction regardless of what we do with the coordinates. All that changes is how we represent that direction mathematically.

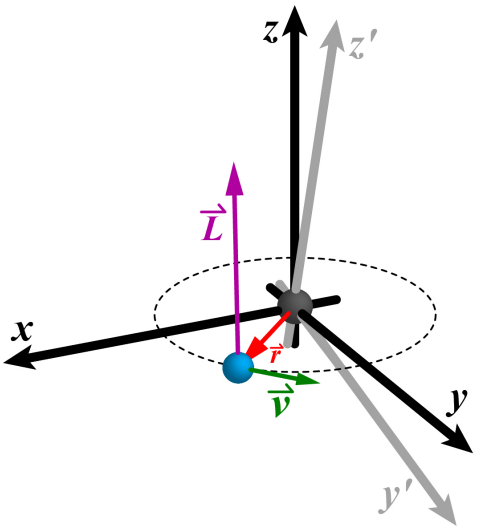


Figure 6.5: The point mass (the solid blue dot) is traveling along the circular path. It's velocity $\vec{v}$, position $\vec{r}$, and angular momentum $\vec{L}$ are given at an arbitrary point along the path.

Unfortunately, even a pseudovector (i.e. a rank-1 pseudotensor) can remain unchanged when a coordinate system is rotated or reflected, so the demonstration given in Figure 6.4 sometimes fails to separate tensors from pseudotensors. However, there has to be some transformation under which they will change otherwise they'd be a real tensor. With pseudovector quantities like angular momentum and torque, translation does the trick for us.

Example 6.5.1

A point mass m is traveling in uniform circular motion with speed v at a distance of R from the origin. Find the angular momentum of this object with the z-axis directed along the axis of rotation. Then, rotate the coordinate system an angle of θ about the x-axis and find the angular momentum again.

- The angular momentum of an object is given by a cross product between the objects position and its linear momentum. In vector equation form, this is

$$\vec{L} = \vec{r} \times \vec{p}$$

where $\vec{r}$ is the position of the object relative to the origin and $\vec{p}$ is the linear momentum of the object. Any quantity defined as a cross product between two *real* vectors is automatically a pseudovector. (Note: If one of the quantities in the cross product is a pseudovector, then the result is a real vector. For example, $\vec{v} = \vec{\omega} \times \vec{r}$ where $\vec{\omega}$ is the pseudovector.)

- If we start with the z-axis as the axis of rotation as shown in Figure 6.5, then we get an angular momentum of

$$\vec{L} = (R\hat{s}) \times \left(mv\hat{\phi}\right) = mvR\hat{z}.$$

- Now we'll do the rotation the easy way by operating a Cartesian rotation matrix on the angular momentum vector. We get

$$\vec{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ mvR \end{bmatrix} = \begin{bmatrix} 0 \\ -mvR\sin\theta \\ mvR\cos\theta \end{bmatrix}$$

$$\boxed{\vec{L} = mvR\left(-\sin\theta\,\hat{y} + \cos\theta\,\hat{z}\right)}$$

which still has a magnitude of mvR. It would appear that the angular momentum has rotated counterclockwise by an angle θ. However, it is really the z-axis which has rotated clockwise. The angular momentum is still directed along the axis of rotation of the point mass and, since its magnitude hasn't changed, we can conclude its fundamental nature hasn't changed either.

Example 6.5.2

A point mass m is traveling in uniform circular motion with speed v at a distance of R from the origin with the z-axis directed along the axis of rotation. Translate the coordinate system by $-R$ along the y-axis and find the angular momentum.

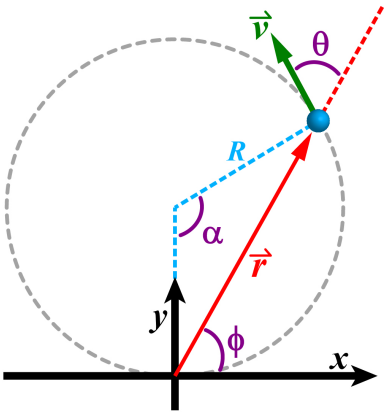


Figure 6.6: The point mass (the solid blue dot) is traveling along the circular (gray dashed) path. It's velocity $\vec{v}$ and position $\vec{r}$ are given at an arbitrary point along the path. A few useful angles are also shown.

- If we shift the coordinates by $-R$ along the y-axis, things get a little tricky. The velocity is still tangent to the path by definition. However, $\vec{r}$ is still defined from the origin to the point mass and now it changes length. It represents a chord of the circle rather than a radius, so we'll have to play some geometry games. Referring to Figure 6.6, we know

$$r = R\,\mathrm{crd}\,\alpha = R\left[2\sin\left(\frac{\alpha}{2}\right)\right] = 2R\sin\left(\frac{\alpha}{2}\right)$$

by the definition of the length of a chord. We also know, by the inscribed angle theorem, that $\theta = \phi$ and $\alpha = 2\phi$, thus

$$r = 2R\sin\phi.$$

- Now that we have r, the angular momentum is

$$\vec{L} = \vec{r}\times\vec{p} = m\left(\vec{r}\times\vec{v}\right) = mvr\sin\phi\,\hat{z}$$

where the factor of $\sin\phi$ comes from Eq. 2.2.3 and we've realized both $\vec{r}$ and $\vec{v}$ are always in the xy-plane. Substituting in for r, we get

$$\vec{L} = mv\left(2R\sin\phi\right)\sin\phi\,\hat{z} = \boxed{2mvR\sin^2\phi\,\hat{z}}.$$

Not only is this not the $mvR\hat{z}$ we got in Example 6.5.1, but it's variable! It changes as the point mass goes around the circle.

- The only way angular momentum can change is if there is an external torque. Torque is defined as

$$\vec{\tau} = \vec{r} \times \vec{F}$$

where $\vec{F}$ is the force causing the curved motion. In this case, it's uniform circular motion, so this force must always point toward the center of the circle (i.e. a centripetal force). Since the angle between $\vec{F}$ and $\vec{v}$ is $\pi/2$ and the angle between $\vec{v}$ and $\vec{r}$ is $\theta = \phi$, we get

$$\vec{\tau} = Fr \sin\Big(\frac{\pi}{2} + \phi\Big)\ \hat{z}.$$

Substituting in what we know of r and centripetal force results in

$$\vec{\tau} = \left(m\frac{v^2}{R}\right)(2R\sin\phi)\sin\Big(\frac{\pi}{2} + \phi\Big)\ \hat{z}$$

$$\vec{\tau} = mv^2\,(2\sin\phi)\sin\Big(\frac{\pi}{2} + \phi\Big)\ \hat{z}.$$

Since $\sin\big(\frac{\pi}{2} + \phi\big) = \cos\phi$, the torque is

$$\vec{\tau} = mv^2\,(2\sin\phi)\cos\phi\,\hat{z} = mv^2\,(2\sin\phi\cos\phi)\,\hat{z}$$

and, since $2\sin\phi\cos\phi = \sin(2\phi)$, our final result is

$$\boxed{\vec{\tau} = mv^2\sin(2\phi)\ \hat{z}},$$

which is also variable. The important point here is the torque in the original coordinate system was zero at all times, yet one little shift of the coordinate system (not the physical system) and suddenly there's a torque. That's the weirdness of pseudotensors. **If real tensors are zero in one coordinate system, they *must* be zero in *all* of them.**

Another way to tell the difference between some tensors and pseudotensors is by changing the physical system. An easy-to-see example is a magnetic field (a pseudovector) generated by a current-carrying wire loop like

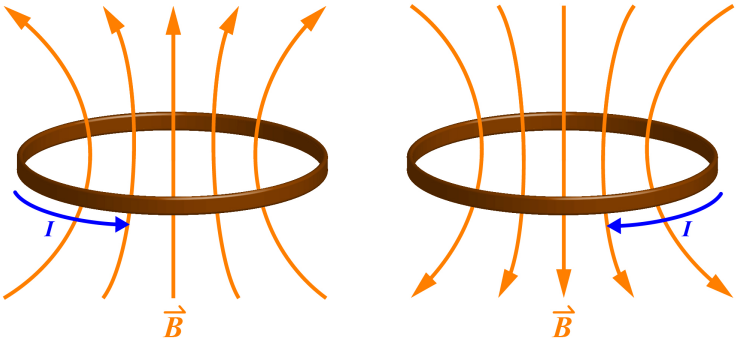


Figure 6.7: On the left, we have a wire loop carrying an electric current in a counterclockwise direction as viewed from above as well as the magnetic field it generates. On the right, we have reflected the scenario on the left horizontally (i.e. across a vertical axis). The direction of the current reflects as we'd expect because its motion is represented by a vector. However, the magnetic field (a pseudovector) gains an extra reflection vertically (i.e. across a horizontal axis).

that shown in Figure 6.7. When the whole scenario is reflected, the magnetic field doesn't reflect in the way you'd expect, but points in the opposite direction. If you're not convinced the B-field is a pseudovector, take a look at the Biot-Savart law (Eq. 5.2.10). It's defined with a cross product of real vectors, which we've already stated makes its status automatic. It turns out that, in general, both $\vec{E}$ and $\vec{B}$ are pseudovectors, but we'll leave that development for a later chapter.

It all really depends on the pseudotensor. Some of them transform just fine under rotations, but not translations (or vice versa). Some of them transform fine between rectilinear coordinates, but not curvilinear. Some of them simply pick up an extra scalar factor when transforming. Others transform in very complex ways. With experience, you just learn which ones are tensors and which are pseudotensors. There's no catch-all rule to figure it out.

6.6 Coordinate Transformations

Typically, in a multi-variable calculus course, we see the use of something called a **Jacobian** to transform between coordinate systems, which works

for many but not all **coordinate transformations**. For example, it doesn't work for the coordinate translation of a position vector, but it will work quite nicely for transforming between the systems described in Chapter 1. The Jacobian that transforms from cylindrical to Cartesian coordinates is

$$J = \begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial \phi} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial \phi} & \frac{\partial y}{\partial z} \\ \frac{\partial z}{\partial s} & \frac{\partial z}{\partial \phi} & \frac{\partial z}{\partial z} \end{bmatrix} = \begin{bmatrix} \cos\phi & -s\sin\phi & 0 \\ \sin\phi & s\cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

which is similar to something we already saw in Eq. 1.2.6. For any dimensional space, we can write this in index notation as

$$J_i^j = \frac{\partial x'^j}{\partial x^i} \tag{6.6.1}$$

which transforms from the unprimed coordinate system to the primed one. When it comes to tensors with multiple indices, each index must be transformed separately. For a contravariant tensor, we have

$$T'^{kl\ldots} = \frac{\partial x'^k}{\partial x^i}\frac{\partial x'^l}{\partial x^j}\cdots T^{ij\ldots} \tag{6.6.2}$$

and, for a covariant tensor, we have

$$T'_{kl\ldots} = \frac{\partial x^i}{\partial x'^k}\frac{\partial x^j}{\partial x'^l}\cdots T_{ij\ldots}, \tag{6.6.3}$$

where the primed coordinates are now on the bottom of the derivative. For a mixed tensor, you simply transform lower indices using the Jacobians found in Eq. 6.6.3 and upper indices using those in Eq. 6.6.2 (Note: Upper indices in the denominator of a derivative are actually lower indices)

Equations involving just tensors are invariant under all coordinate transformations because the transformations are just multiplicative factors which will cancel on either side. Pseudotensors, on the other hand, do not always transform according to Eqs. 6.6.2 and/or 6.6.3. This makes equations involving them a challenge at times. However, if the transformation doesn't vary much from that of a tensor, then it isn't too difficult to adjust.

Example 6.6.1

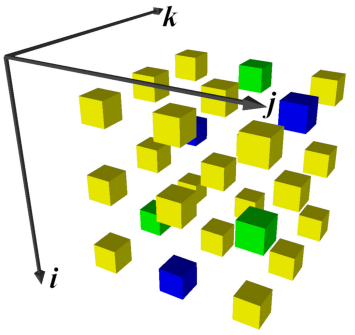


Figure 6.8: This is the rank-3 Levi-Civita pseudotensor, ε_{ijk}, in matrix notation. Yellow boxes represent a zero, green a 1, and blue a -1. It is clear only 6 of the $3^3 = 27$ components are non-zero.

The angular momentum of an object is given by a cross product between the objects position and its linear momentum. In vector equation form, this is

$$\vec{L} = \vec{r} \times \vec{p}$$

where $\vec{r}$ is the position of the object relative to the origin and $\vec{p}$ is the linear momentum of the object. If we want to write any cross product in index notation, then we need to use a special rank-3 pseudotensor called the **Levi-Civita pseudotensor**,

$$\varepsilon_{ijk} = \begin{cases} +1, & \text{if } (i,j,k) \text{ is an even permutation of } (1,2,3) \\ -1, & \text{if } (i,j,k) \text{ is an odd permutation of } (1,2,3) \\ 0, & \text{otherwise} \end{cases} \tag{6.6.4}$$

where i, j, and k can each take on the. It's special in that it's antisymmetric (i.e. $T_{ij} = -T_{ji}$) and also unit (i.e. composed of unit and/or zero vector sections, but is not the zero-tensor). Using this, we can write the angular momentum as

$$L_k = \varepsilon_{ijk} r^i p^j$$

in the coordinate basis or

$$L_{\hat{k}} = \varepsilon_{\hat{i}\hat{j}\hat{k}} r^{\hat{i}} p^{\hat{j}}.$$

in the orthonormal basis, which is probably the more familiar for most of us.

- For example, let's say we have a point mass m traveling in uniform circular motion with speed v at a distance of R from the origin with the z-axis directed along the axis of rotation. For the sake of simplicity, we'll work in cylindrical coordinates starting in the orthonormal basis $\left\{\hat{s}, \hat{\phi}, \hat{z}\right\}$. Under these circumstances, the angular momentum is

$$L_{\hat{k}} = \varepsilon_{\hat{s}\hat{j}\hat{k}} r^{\hat{s}} p^{\hat{j}} + \varepsilon_{\hat{\phi}\hat{j}\hat{k}} r^{\hat{\phi}} p^{\hat{j}} + \varepsilon_{\hat{z}\hat{j}\hat{k}} r^{\hat{z}} p^{\hat{j}}$$

$$\begin{aligned} L_{\hat{k}} &= \varepsilon_{\hat{s}\hat{s}\hat{k}} r^{\hat{s}} p^{\hat{s}} + \varepsilon_{\hat{s}\hat{\phi}\hat{k}} r^{\hat{s}} p^{\hat{\phi}} + \varepsilon_{\hat{s}\hat{z}\hat{k}} r^{\hat{s}} p^{\hat{z}} \\ &\quad + \varepsilon_{\hat{\phi}\hat{s}\hat{k}} r^{\hat{\phi}} p^{\hat{s}} + \varepsilon_{\hat{\phi}\hat{\phi}\hat{k}} r^{\hat{\phi}} p^{\hat{\phi}} + \varepsilon_{\hat{\phi}\hat{z}\hat{k}} r^{\hat{\phi}} p^{\hat{z}} \\ &\quad + \varepsilon_{\hat{z}\hat{s}\hat{k}} r^{\hat{z}} p^{\hat{s}} + \varepsilon_{\hat{z}\hat{\phi}\hat{k}} r^{\hat{z}} p^{\hat{\phi}} + \varepsilon_{\hat{z}\hat{z}\hat{k}} r^{\hat{z}} p^{\hat{z}} \end{aligned}$$

having expanded over both sums (i.e. both i and j). By Eq. 6.6.4, this simplifies to

$$\begin{aligned} L_{\hat{k}} &= \varepsilon_{\hat{s}\hat{\phi}\hat{k}} r^{\hat{s}} p^{\hat{\phi}} + \varepsilon_{\hat{s}\hat{z}\hat{k}} r^{\hat{s}} p^{\hat{z}} \\ &\quad + \varepsilon_{\hat{\phi}\hat{s}\hat{k}} r^{\hat{\phi}} p^{\hat{s}} + \varepsilon_{\hat{\phi}\hat{z}\hat{k}} r^{\hat{\phi}} p^{\hat{z}} \\ &\quad + \varepsilon_{\hat{z}\hat{s}\hat{k}} r^{\hat{z}} p^{\hat{s}} + \varepsilon_{\hat{z}\hat{\phi}\hat{k}} r^{\hat{z}} p^{\hat{\phi}}, \end{aligned}$$

where k can still take on any value. We can write out the three components separately while using Eq. 6.6.4 again to get

$$\left\{\begin{array}{rcl} L_{\hat{s}} &=& \varepsilon_{\hat{\phi}\hat{z}\hat{s}} r^{\hat{\phi}} p^{\hat{z}} + \varepsilon_{\hat{z}\hat{\phi}\hat{s}} r^{\hat{z}} p^{\hat{\phi}} \\ L_{\hat{\phi}} &=& \varepsilon_{\hat{s}\hat{z}\hat{\phi}} r^{\hat{s}} p^{\hat{z}} + \varepsilon_{\hat{z}\hat{s}\hat{\phi}} r^{\hat{z}} p^{\hat{s}} \\ L_{\hat{z}} &=& \varepsilon_{\hat{s}\hat{\phi}\hat{z}} r^{\hat{s}} p^{\hat{\phi}} + \varepsilon_{\hat{\phi}\hat{s}\hat{z}} r^{\hat{\phi}} p^{\hat{s}} \end{array}\right\}$$

$$\left\{\begin{array}{rcl} L_{\hat{s}} &=& r^{\hat{\phi}} p^{\hat{z}} - r^{\hat{z}} p^{\hat{\phi}} \\ L_{\hat{\phi}} &=& -r^{\hat{s}} p^{\hat{z}} + r^{\hat{z}} p^{\hat{s}} \\ L_{\hat{z}} &=& r^{\hat{s}} p^{\hat{\phi}} - r^{\hat{\phi}} p^{\hat{s}} \end{array}\right\},$$

which is exactly what you'd expect for the components of a cross product. We also know $\vec{r} = R\hat{s}$ and $\vec{p} = m\vec{v} = mv\hat{\phi}$, which makes everything disappear except the first term in the z-component. The angular momentum is

$$L_{\hat{z}} = r^{\hat{s}} p^{\hat{\phi}} = (R)\,(mv) = mvR,$$

which is exactly what we expected, so no problems there. It might not be the most efficient way to solve the problem, but at least it shows consistency.

- So what happens in the coordinate basis? It's almost the same process, except based on Eq. 3.4.1, we have

$$\left.\begin{aligned}\vec{r} &= R\hat{s} = R\vec{e}_s \\ \vec{p} &= m\vec{v} = mv\hat{\phi} = \frac{mv}{R}R\hat{\phi} = \frac{mv}{R}\vec{e}_\phi\end{aligned}\right\},$$

which makes linear momentum look a little strange. That's what we get for using a coordinate basis. The resulting angular momentum is

$$L_z = r^s p^\phi = (R)\left(\frac{mv}{R}\right) = mv,$$

which doesn't make much sense. Linear momentum changed its appearance because $\hat{\phi} \neq \vec{e}_\phi$, so it might not be too surprising at this point. However, we know $\hat{z} = \vec{e}_z$ because $h_z = 1$ (see Section 3.4), so it shouldn't be any different (i.e. $L_z = L_{\hat{z}}$). What the heck happened?! How did we lose a factor of R?

This actually comes down to the fact that the pseudotensor ε_{ijk} is what makes angular momentum (and every other result of a cross product) a pseudovector. The Levi-Civita pseudotensor transforms by

$$\varepsilon'_{lmn} = \frac{\partial x^i}{\partial x'^l}\frac{\partial x^j}{\partial x'^m}\frac{\partial x^k}{\partial x'^n}\varepsilon_{ijk}\ \det(J)\,.$$

which looks a lot like Eq. 6.6.3 with an extra factor of $\det(J)$. If the primed system is Cartesian, then $\det(J) = \sqrt{|\det(g)|}$, where g is the metric tensor of the space. We can now write the transformation as

$$\boxed{\varepsilon'_{lmn} = \frac{\partial x^i}{\partial x'^l}\frac{\partial x^j}{\partial x'^m}\frac{\partial x^k}{\partial x'^n}\varepsilon_{ijk}\sqrt{|\det(g)|}}\,. \tag{6.6.5}$$

You might be thinking "Hey! We transformed from a *cylindrical* orthonormal basis, not from a *Cartesian* orthonormal basis!" Well, Eq. 6.4.9 says the orthonormal metric is equivalent to the Cartesian metric regardless of your system. It's subtle, but it works in our favor.

With Eq. 6.6.5 in mind, the equation for angular momentum is actually

$$\boxed{L_k = \sqrt{|\det(g)|}\,\varepsilon_{ijk} r^i p^j}, \tag{6.6.6}$$

which applies to both a coordinate and an orthonormal basis (since $|\det(g)| = 1$ in the orthonormal basis). Since

$$g_{ij} \longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & s^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6.6.7}$$

means $\sqrt{|\det(g)|} = s$ and we know $s = R$ at all times for our point mass, we arrive once again at $L_z = mvR$.

6.7 Tensor Calculus

With the exception of a few differential coordinates, all we've seen so far is tensor algebra. However, most physics is about changes, so eventually we'll have to take a derivative of a tensor. The procedure for doing so can be rather complicated depending on the chosen coordinate system. In Cartesian coordinates, it isn't so bad. All we have to do is operate with the del operator (Eq. 3.2.1), which can be written index notation as

$$\nabla_i T^j = \frac{\partial}{\partial x^i} T^j = \frac{\partial T^j}{\partial x^i}$$

for vectors (rank-1 tensors) or

$$\nabla_i T^{jk} = \frac{\partial}{\partial x^i} T^{jk} = \frac{\partial T^{jk}}{\partial x^i}$$

for rank-2 tensors (Note: Upper indices in the denominator of a derivative are actually lower indices). Piece of cake, right? Well, not quite. These only look simple and familiar because of the nature of Cartesian space. We know from Section 3.3 the del operator isn't always so simple.

In general, we need to be careful. Let's recall the definition of a derivative from single-variable calculus:

$$\frac{df}{dx} \equiv \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}. \tag{6.7.1}$$

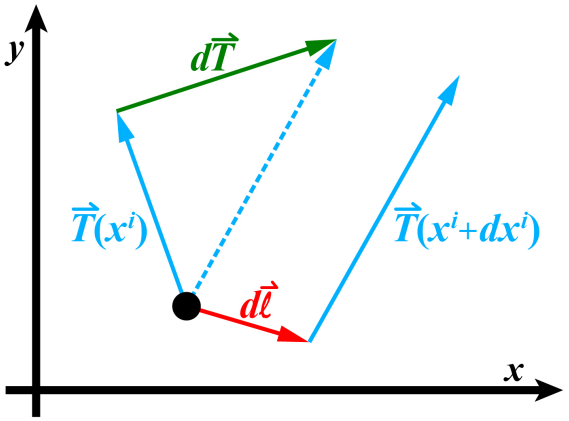


Figure 6.9: This is a demonstration of the parallel transport of a vector T^i. The dashed blue vector represents the vector $\vec{T}(x^i + dx^i)$ at x^i. This move was necessary to subtract $\vec{T}(x^i)$ from $\vec{T}(x^i + dx^i)$.

It is often misinterpreted that, ultimately, $\Delta x = 0$ and the limit is just a way to get there without violating fundamental mathematics. In actuality, Δx is *never* zero, it just *approaches* it becoming dx. Sure, it gets pretty close. So close, in fact, that we can approximate it that way (the ultimate power of the limit). However, it can't be exactly zero because it's in the denominator of a fraction.

The point here is that Eq. 6.7.1 is always discussing two distinct values: x and $x + dx$. If we extend this concept to 3-space (or just 2-space for that matter), then our issue becomes clear. The numerator of Eq. 6.7.1 involves a subtraction of functions, which in 3-space would be tensor functions. For the simplicity of our discussion, we'll assume the tensor is just a vector (we'll discuss higher orders later). In order to subtract vectors properly, we need them to be at *exactly* the same place (i.e. not separated by $d\vec{\ell}$). In short, we have to move one of them and that's where things get tricky.

The process of moving a vector in space for addition or subtraction is called **parallel transport** (See Figure 6.9). We have to make sure the vector at its new location is parallel to itself at the old location to guarantee it's still the same vector. In Cartesian coordinates, a vector translation isn't going to change the vector, so we get what we expected at the beginning of this section. However, in a curvilinear coordinate system, it's a completely different story.

If a vector changes in magnitude and/or direction when it's translated, then we need to have some kind of adjustment for it. This is only really important when taking a derivative, so we'll just adjust the derivative. This involves a pseudotensor quantity known as a **Christoffel symbol**, Γ, defined by

$$\nabla_i \left(\vec{e}_l\right)_j \equiv \Gamma^k_{ij} \left(\vec{e}_l\right)_k , \tag{6.7.2}$$

in the coordinate basis since we can place the blame entirely on the basis vector. We know it's a pseudotensor because it's the zero-tensor in some coordinates, but non-zero in others. It is also symmetric over the lower two indices (i.e. $\Gamma^k_{ij} = \Gamma^k_{ji}$). This means the del operation (sometimes called the **covariant derivative**) is actually given by

$$\nabla_i T^j = \frac{\partial T^j}{\partial x^i} + \Gamma^j_{ik} T^k \tag{6.7.3}$$

for a contravariant vector. The Christoffel term represents our small shift in the vector's position for the derivative and is, by no means, insignificant. For a covariant vector, we get

$$\nabla_i T_j = \frac{\partial T_j}{\partial x^i} - \Gamma^k_{ij} T_k , \tag{6.7.4}$$

where we've swapped the indices and the sign of the extra term to compensate for the change.

This can work for higher rank tensors as well, but we need a Christoffel term for each tensor index. For a contravariant rank-2 tensor, this is

$$\nabla_i T^{jk} = \frac{\partial T^{jk}}{\partial x^i} + \Gamma^j_{il} T^{lk} + \Gamma^k_{il} T^{jl} , \tag{6.7.5}$$

where first Christoffel term sums over the first index on T^{jk} (i.e. it adjusts the derivative for the first index) and the second Christoffel terms sums over the second index (i.e. it adjusts the derivative for the second index). The appropriate Christoffel term in Eq. 6.7.5 can be changed as they were in Eq. 6.7.4 to account a covariant index.

Now we're only left with one question: "How do we find the Christoffel symbols for a given space?" Any adjustment we make to a tensor when we

move it in a coordinate system is going to be related to how that coordinate system changes in space. We've already learned the metric tensor is what describes the space, so the Christoffel symbols should be related to changes in the metric tensor. The relationship is

$$\Gamma^k_{ij} = \frac{1}{2} g^{lk} \left(\frac{\partial g_{li}}{\partial x^j} + \frac{\partial g_{lj}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^l} \right), \tag{6.7.6}$$

where i, j, and k are free indices unlike l which is a summation index. We should also know that Eq. 6.7.6 (the origin of which we'll explain later) involves both the metric tensor, g_{ij}, and its inverse g^{ij}. You'll need both to find the Christoffel symbols.

Example 6.7.1

Find the Christoffel symbols in a set of arbitrary orthogonal coordinates, (q^1, q^2, q^3).

- First, we need to know the metric tensor for the space. If the coordinate basis vectors are orthogonal, then Eq. 6.4.4 tells us the metric tensor is diagonal taking the form

$$g_{ij} \longrightarrow \begin{bmatrix} h_1 h_1 & 0 & 0 \\ 0 & h_2 h_2 & 0 \\ 0 & 0 & h_3 h_3 \end{bmatrix}, \tag{6.7.7}$$

  where we've avoided using exponents for reasons that should become clear as we go through the solution. This makes it's inverse

$$g^{ij} \longrightarrow \begin{bmatrix} \frac{1}{h_1 h_1} & 0 & 0 \\ 0 & \frac{1}{h_2 h_2} & 0 \\ 0 & 0 & \frac{1}{h_3 h_3} \end{bmatrix}. \tag{6.7.8}$$

- There are $3^3 = 27$ Christoffel symbols in total and we'll be using Eq. 6.7.6 to find them. We have to be careful with the Einstein summation convention, but we should still be able to shorten our work by taking advantage of the diagonal nature of the metric tensor and the symmetric in the Christoffel symbol.

- If $i = j = k = 1$, then we get

$$\Gamma^1_{11} = \frac{1}{2} g^{l1} \left(\frac{\partial g_{l1}}{\partial q^1} + \frac{\partial g_{l1}}{\partial q^1} - \frac{\partial g_{11}}{\partial q^l} \right).$$

We still have a summation over l, so there are actually 3 giant terms that take the above form. However, we know g^{ij} is diagonal, so the only non-zero term is $l = 1$. We now get

$$\Gamma^1_{11} = \frac{1}{2} g^{11} \left(\frac{\partial g_{11}}{\partial q^1} + \frac{\partial g_{11}}{\partial q^1} - \frac{\partial g_{11}}{\partial q^1} \right) = \frac{1}{2} g^{11} \frac{\partial g_{11}}{\partial q^1}$$

Now we can substitute in the components of the metric and its inverse to get

$$\Gamma^1_{11} = \frac{1}{2} \frac{1}{h_1 h_1} \frac{\partial}{\partial q^1} (h_1 h_1)$$

We can use Eq. 4.2.8 to simplify and also do this same process for the other two values of $i = j = k$, which gives us

$$\boxed{\left\{ \begin{array}{rcl} \Gamma^1_{11} & = & \dfrac{1}{h_1} \dfrac{\partial h_1}{\partial q^1} \\ \Gamma^2_{22} & = & \dfrac{1}{h_2} \dfrac{\partial h_2}{\partial q^2} \\ \Gamma^3_{33} & = & \dfrac{1}{h_3} \dfrac{\partial h_3}{\partial q^3} \end{array} \right\}}.$$

That's three Christoffel symbols so far.

- If $i = k = 1$ and $j = 2$, then we get

$$\Gamma^1_{12} = \frac{1}{2} g^{l1} \left(\frac{\partial g_{l1}}{\partial q^2} + \frac{\partial g_{l2}}{\partial q^1} - \frac{\partial g_{12}}{\partial q^l} \right).$$

We still have a summation over l, so there are actually 3 giant terms that take the above form. However, we know g^{ij} is diagonal, so the only non-zero term is $l = 1$. We now get

$$\Gamma^1_{12} = \frac{1}{2} g^{11} \left(\frac{\partial g_{11}}{\partial q^2} + \frac{\partial g_{12}}{\partial q^1} - \frac{\partial g_{12}}{\partial q^1} \right).$$

Since g_{ij} is also diagonal, the last two terms in parentheses are zero. Now we can substitute in the components of the metric and its inverse to get

$$\Gamma^1_{12} = \frac{1}{2} g^{11} \frac{\partial g_{11}}{\partial q^2} = \frac{1}{2} \frac{1}{h_1 h_1} \frac{\partial}{\partial q^2} \left(h_1 h_1 \right).$$

We can use Eq. 4.2.8 to simplify and also do this same process for similar index patterns, which gives us

$$\boxed{\left\{ \begin{array}{l} \Gamma^1_{12} = \Gamma^1_{21} = \dfrac{1}{h_1} \dfrac{\partial h_1}{\partial q^2} \\ \Gamma^1_{13} = \Gamma^1_{31} = \dfrac{1}{h_1} \dfrac{\partial h_1}{\partial q^3} \\ \Gamma^2_{21} = \Gamma^2_{12} = \dfrac{1}{h_2} \dfrac{\partial h_2}{\partial q^1} \\ \text{etc.} \end{array} \right\},}$$

noting that the Christoffel symbol is symmetric over the bottom two indices. That's 12 more Christoffel symbols for a total of 15.

- If $i = j = 1$ (i.e. lower two indices are the same) and $k = 2$, then we get

$$\Gamma^2_{11} = \frac{1}{2} g^{l2} \left(\frac{\partial g_{l1}}{\partial q^1} + \frac{\partial g_{l1}}{\partial q^1} - \frac{\partial g_{11}}{\partial q^l} \right).$$

We still have a summation over l, so there are actually 3 giant terms that take the above form. However, we know g^{ij} is diagonal, so the only non-zero term is $l = 2$. We now get

$$\Gamma^2_{11} = \frac{1}{2} g^{22} \left(\frac{\partial g_{21}}{\partial q^1} + \frac{\partial g_{21}}{\partial q^1} - \frac{\partial g_{11}}{\partial q^2} \right).$$

Since g_{ij} is also diagonal, the first two terms in parentheses are zero. Now we can substitute in the components of the metric and its inverse to get

$$\Gamma^2_{11} = -\frac{1}{2} g^{22} \frac{\partial g_{11}}{\partial q^2} = -\frac{1}{2} \frac{1}{h_2 h_2} \frac{\partial}{\partial q^2} \left(h_1 h_1 \right).$$

We can use Eq. 4.2.8 to simplify and also do this same process for similar index patterns, which gives us

$$\boxed{\begin{cases} \Gamma^2_{11} = -\dfrac{h_1}{h_2 h_2}\dfrac{\partial h_1}{\partial q^2} \\ \Gamma^3_{11} = -\dfrac{h_1}{h_3 h_3}\dfrac{\partial h_1}{\partial q^3} \\ \Gamma^1_{22} = -\dfrac{h_2}{h_1 h_1}\dfrac{\partial h_2}{\partial q^1} \\ \text{etc.} \end{cases}}.$$

That's six more Christoffel symbols for a total of 21.

- We only need six more to make 27 and they correspond to when i, j, and k are all different. If $(i, j, k) = (1, 2, 3)$, then we get

$$\Gamma^3_{12} = \frac{1}{2} g^{l3}\left(\frac{\partial g_{l1}}{\partial q^2} + \frac{\partial g_{l2}}{\partial q^1} - \frac{\partial g_{12}}{\partial q^l}\right).$$

We still have a summation over l, so there are actually 3 giant terms that take the above form. However, we know g^{ij} is diagonal, so the only non-zero term is $l = 3$. We now get

$$\Gamma^3_{12} = \frac{1}{2} g^{33}\left(\frac{\partial g_{31}}{\partial q^2} + \frac{\partial g_{32}}{\partial q^1} - \frac{\partial g_{12}}{\partial q^3}\right).$$

Since g_{ij} is also diagonal, the entire Christoffel symbol is zero. This occurs will all the remaining symbols which we can state as

$$\boxed{\Gamma^3_{12} = \Gamma^3_{21} = \Gamma^1_{23} = \Gamma^1_{32} = \Gamma^2_{13} = \Gamma^2_{31} = 0}.$$

That's a total of 27 Christoffel symbols!

Example 6.7.2

Use tensor analysis to find the divergence of a vector, A^j, in a set of arbitrary orthogonal coordinates, (q^1, q^2, q^3).

- The divergence of a vector is a covariant derivative as given by Eq. 6.7.3. However, Eq. 6.7.3 as it stands has two free indices, which results in a rank-2 tensor. A vector divergence always results in a scalar, so we need no free indices in our result. If a summation results in a scalar, it is referred to as a **scalar product** (i.e. a generalized dot product). That in mind, we can now say

$$\nabla_i A^i = \frac{\partial A^i}{\partial q^i} + \Gamma^i_{ik} A^k, \tag{6.7.9}$$

where i is now a summation index and there are no free indices. The index k also represents its own summation independent from i. If we expand both summations, then we have

$$\nabla_i A^i = \nabla_1 A^1 + \nabla_2 A^2 + \nabla_3 A^3,$$

where

$$\left\{ \begin{array}{l} \nabla_1 A^1 = \dfrac{\partial A^1}{\partial q^1} + \Gamma^1_{11} A^1 + \Gamma^1_{12} A^2 + \Gamma^1_{13} A^3 \\ \nabla_2 A^2 = \dfrac{\partial A^2}{\partial q^2} + \Gamma^2_{21} A^1 + \Gamma^2_{22} A^2 + \Gamma^2_{23} A^3 \\ \nabla_3 A^3 = \dfrac{\partial A^3}{\partial q^3} + \Gamma^3_{31} A^1 + \Gamma^3_{32} A^2 + \Gamma^3_{33} A^3 \end{array} \right\}$$

- These are all added together anyway, so let's consider just the A^1 terms for now. Using the Christoffel symbols we found in Example 6.7.1, we get

$$\frac{\partial A^1}{\partial q^1} + \Gamma^1_{11} A^1 + \Gamma^2_{21} A^1 + \Gamma^3_{31} A^1$$

$$\frac{\partial A^1}{\partial q^1} + \frac{1}{h_1}\frac{\partial h_1}{\partial q^1} A^1 + \frac{1}{h_2}\frac{\partial h_2}{\partial q^1} A^1 + \frac{1}{h_3}\frac{\partial h_3}{\partial q^1} A^1$$

$$\frac{1}{h_1 h_2 h_3}\left[h_1 h_2 h_3 \frac{\partial A^1}{\partial q^1} + h_2 h_3 \frac{\partial h_1}{\partial q^1} A^1 + h_1 h_3 \frac{\partial h_2}{\partial q^1} A^1 + h_1 h_2 \frac{\partial h_3}{\partial q^1} A^1 \right].$$

The quantity in brackets just looks like one big derivative product rule (defined by Eq. 3.1.5), so we can simplify this drastically by saying

$$\frac{1}{h_1 h_2 h_3}\frac{\partial}{\partial q^1}\left(h_1 h_2 h_3 A^1\right)$$

This may not look familiar since we're working in the coordinate basis. Using Eq. 3.4.1, we can say $A^1\vec{e}_1 = A^1 h_1 \hat{e}_1 = A^{\hat{1}}\hat{e}_1$ means $A^1 h_1 = A^{\hat{1}}$. This leaves us with

$$\frac{1}{h_1 h_2 h_3}\frac{\partial}{\partial q^1}\left(h_2 h_3 A^{\hat{1}}\right)$$

- A very similar process happens with the A^2 and A^3 terms. The total result is

$$\boxed{\nabla_i A^i = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial q^1}\left(h_2 h_3 A^{\hat{1}}\right) + \frac{\partial}{\partial q^1}\left(h_1 h_3 A^{\hat{2}}\right) + \frac{\partial}{\partial q^1}\left(h_1 h_2 A^{\hat{3}}\right)\right]}$$

which is exactly Eq. 3.4.9.

Example 6.7.3

Use tensor analysis to find the curl of a vector, A^j, in a set of arbitrary orthogonal coordinates, (q^1, q^2, q^3).

- The curl of a vector is a bit more complicated than the divergence because it involves the cross product. We have some experience with this from Example 6.6.1 where we defined the angular momentum by Eq. 6.6.6. Similarly, the curl of a vector can be written as

$$\left(\vec{\nabla}\times\vec{A}\right)_m = \sqrt{|\det(g)|}\,\varepsilon_{mkj}\nabla^k A^j.$$

However, a contravariant derivative isn't really convenient. We can use $\nabla^k = g^{ki}\nabla_i$ (a process described in Section 6.4) to make it a covariant derivative, resulting in

$$\left(\vec{\nabla}\times\vec{A}\right)_m = \sqrt{|\det(g)|}\,\varepsilon_{mkj}g^{ki}\nabla_i A^j.$$

We'd also like to raise the index on the left side so that we're dealing with contravariant vector components (because they're easier to picture). Operating with the inverse metric, g^{lm}, the result is

$$\left(\vec{\nabla} \times \vec{A}\right)^l = \sqrt{|\det(g)|}\, g^{lm} \varepsilon_{mkj} g^{ki} \nabla_i A^j \tag{6.7.10}$$

where ε_{mkj} is the Levi-Civita pseudotensor described by Eq. 6.6.4. We also know from Eq. 6.7.7 that $\sqrt{|\det(g)|} = h_1 h_2 h_3$, so

$$\left(\vec{\nabla} \times \vec{A}\right)^l = h_1 h_2 h_3\, g^{lm} \varepsilon_{mkj} g^{ki} \nabla_i A^j.$$

- Let's start by considering the first component of the curl. This is given by

$$\left(\vec{\nabla} \times \vec{A}\right)^1 = h_1 h_2 h_3\, g^{1m} \varepsilon_{mkj} g^{ki} \nabla_i A^j.$$

There is a summation over m, so there are actually 3 giant terms that take the above form. However, we know g^{lm} is diagonal, so the only non-zero term is $m = 1$. We now get

$$\left(\vec{\nabla} \times \vec{A}\right)^1 = h_1 h_2 h_3\, g^{11} \varepsilon_{1kj} g^{ki} \nabla_i A^j$$

$$\left(\vec{\nabla} \times \vec{A}\right)^1 = \frac{h_2 h_3}{h_1} \varepsilon_{1kj} g^{ki} \nabla_i A^j$$

where we've made a substitution from Eq. 6.7.8.

- There are two other summations over indices k and j. According to Eq. 6.6.4, the indices of the Levi-Civita pseudotensor must all be different for a non-zero value. Since we already know $m = 1$, we know that $kj = 23$ and $kj = 32$ are the only non-zero terms. The result is

$$\left(\vec{\nabla} \times \vec{A}\right)^1 = \frac{h_2 h_3}{h_1} \left[\varepsilon_{123} g^{2i} \nabla_i A^3 + \varepsilon_{132} g^{3i} \nabla_i A^2\right]$$

$$\left(\vec{\nabla} \times \vec{A}\right)^1 = \frac{h_2 h_3}{h_1} \left[g^{2i} \nabla_i A^3 - g^{3i} \nabla_i A^2\right].$$

Again, g^{ki} is diagonal, so the only non-zero terms in the sum over i are

$$\left(\vec{\nabla}\times\vec{A}\right)^1 = \frac{h_2h_3}{h_1}\left[g^{22}\nabla_2A^3 - g^{33}\nabla_3A^2\right].$$

$$\left(\vec{\nabla}\times\vec{A}\right)^1 = \frac{h_2h_3}{h_1}\left[\frac{1}{h_2h_2}\nabla_2A^3 - \frac{1}{h_3h_3}\nabla_3A^2\right].$$

where we've made a substitution from Eq. 6.7.8. We can simplify further to

$$\left(\vec{\nabla}\times\vec{A}\right)^1 = \frac{1}{h_1h_2h_3}\left[h_3h_3\nabla_2A^3 - h_2h_2\nabla_3A^2\right] \tag{6.7.11}$$

- The two terms in brackets are defined by Eq. 6.7.3. They are

$$\left\{\begin{array}{l}\nabla_2A^3 = \dfrac{\partial A^3}{\partial q^2} + \Gamma^3_{2n}A^n \\ \nabla_3A^2 = \dfrac{\partial A^2}{\partial q^3} + \Gamma^2_{3n}A^n\end{array}\right\}$$

$$\left\{\begin{array}{l}\nabla_2A^3 = \dfrac{\partial A^3}{\partial q^2} + \Gamma^3_{21}A^1 + \Gamma^3_{22}A^2 + \Gamma^3_{23}A^3 \\ \nabla_3A^2 = \dfrac{\partial A^2}{\partial q^3} + \Gamma^2_{31}A^1 + \Gamma^2_{32}A^2 + \Gamma^2_{33}A^3\end{array}\right\}.$$

Using the Christoffel symbols we found in Example 6.7.1, we get

$$\left\{\begin{array}{l}\nabla_2A^3 = \dfrac{\partial A^3}{\partial q^2} - \dfrac{h_2}{h_3h_3}\dfrac{\partial h_2}{\partial q^3}A^2 + \dfrac{1}{h_3}\dfrac{\partial h_3}{\partial q^2}A^3 \\ \nabla_3A^2 = \dfrac{\partial A^2}{\partial q^3} + \dfrac{1}{h_2}\dfrac{\partial h_2}{\partial q^3}A^2 - \dfrac{h_3}{h_2h_2}\dfrac{\partial h_3}{\partial q^2}A^3\end{array}\right\}.$$

- These are both added together with their respective coefficients, so let's consider just the A^3 terms for now. This would be

$$h_3h_3\left[\frac{\partial A^3}{\partial q^2} + \frac{1}{h_3}\frac{\partial h_3}{\partial q^2}A^3\right] - h_2h_2\left[-\frac{h_3}{h_2h_2}\frac{\partial h_3}{\partial q^2}A^3\right]$$

$$h_3 h_3 \frac{\partial A^3}{\partial q^2} + h_3 \frac{\partial h_3}{\partial q^2} A^3 + h_3 \frac{\partial h_3}{\partial q^2} A^3$$

$$h_3 h_3 \frac{\partial A^3}{\partial q^2} + 2 h_3 \frac{\partial h_3}{\partial q^2} A^3.$$

We can use Eq. 4.2.8 on the second term to get

$$h_3 h_3 \frac{\partial A^3}{\partial q^2} + \frac{\partial}{\partial q^2} \left(h_3 h_3 \right) A^3,$$

which is just the derivative product rule (defined by Eq. 3.1.5). Simplifying further, we arrive at

$$\frac{\partial}{\partial q^2} \left(h_3 h_3 A^3 \right)$$

- A similar process can be done on the A^2 terms and we can substitute both back into Eq. 6.7.11. The result is

$$\left(\vec{\nabla} \times \vec{A} \right)^1 = \frac{1}{h_1 h_2 h_3} \left[\frac{\partial}{\partial q^2} \left(h_3 h_3 A^3 \right) - \frac{\partial}{\partial q^3} \left(h_2 h_2 A^2 \right) \right],$$

which may look unfamiliar since we're working in the coordinate basis. Using Eq. 3.4.1, we can say $A^2 \vec{e}_2 = A^2 h_2 \hat{e}_2 = A^{\hat{2}} \hat{e}_2$ means $A^2 h_2 = A^{\hat{2}}$ (and similarly for A^3). This leaves us with

$$\left(\vec{\nabla} \times \vec{A} \right)^1 = \frac{1}{h_1 h_2 h_3} \left[\frac{\partial}{\partial q^2} \left(h_3 A^{\hat{3}} \right) - \frac{\partial}{\partial q^3} \left(h_2 A^{\hat{2}} \right) \right],$$

but we also have move to the orthonormal basis on the left side as well. If $C^{\hat{1}} = C^1 h_1$, then

$$\boxed{\left(\vec{\nabla} \times \vec{A} \right)^{\hat{1}} = \frac{1}{h_2 h_3} \left[\frac{\partial}{\partial q^2} \left(h_3 A^{\hat{3}} \right) - \frac{\partial}{\partial q^3} \left(h_2 A^{\hat{2}} \right) \right]},$$

which is exactly the $\hat{e}_1$ component in Eq. 3.4.10. The other two components follow the same pattern.

Chapter 7

Special Relativity

7.1 Origins

Since the early-to-middle 17^{th} century, we've been keenly aware that motion is relative. Let's say you're an baseball outfielder. If you throw the baseball at 30 mph toward the second base while running at 15 mph toward second base, then the player at second base is going to see the ball approaching them at 45 mph. Each person their own perspective known as a **frame of reference**. The concept is often called "classical relativity" or sometimes "Galilean relativity" because it was Galileo who first formalized it.

However, in the late 19^{th} century, the field of electrodynamics had developed into a very solid theory (See Chapter 5) and with it came a very big problem. From Eq. 5.5.2, we discovered the speed of light, c, was constant (defined by Eq. 5.5.4). There is no indication of any dependence on time, space, or perspective. It is a universal constant and it is finite.

Let's take another look at our baseball example. You're running again at 30 mph toward second base, but this time you're pointing a flashlight rather than throwing a ball. According to classical relativity, the player at second base should see the *light* approaching at $c + 30$ mph. Mind you, c is a little under 671 million mph, so 30 mph more isn't going to change it much. Fundamentally though, this is still a problem because it still changes the speed of light regardless of how little. According to electrodynamics, the speed of light is not dependent on perspective, so the second-base player should still see the light approaching at *exactly* c. There in lies our problem.

It was widely accepted that neither classical relativity nor electrodynam-

ics could be drastically wrong. Since classical relativity was the least abstract and easiest to test, it was believed the problem lied with electrodynamics in some minor way. It was suggested that maybe Maxwell's equations (Eq. Set 5.4.9) are defined in the rest frame of the medium in which light propagates (what they called luminiferous aether), so c only takes on the value given by Eq. 5.5.4 in that frame of reference. It was then a mission for physics to find out how the aether was moving relative to the Earth.

Many optical experiments were done in the effort (the most famous of which by Albert Michelson and Edward Morley in 1887). None of the experiments succeeded in measuring the velocity of the aether, which leaves us with only four possible conclusions:

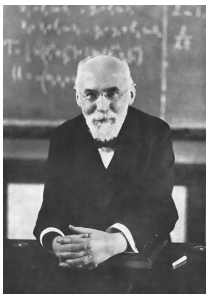
1. *The Earth is in the rest frame of the aether.* This is highly unlikely since the Earth travels in an ellipse (nearly a circle) around the sun. The Earth's motion is continually changing direction, so this can't be true all the time.

2. *The Earth carries a pocket of aether with it as it moves.* This is akin to what we'd see around a car in a wind tunnel. The car forms a pocket of stationary air (relative to the car) around itself as it moves, which is why bugs can land on your windshield while your car is stationary and stay there for the whole trip with little effort. Applying this conclusion to the luminiferous aether was very popular at the time, but unsubstantiated by other evidence.

3. *The aether had the power to contract the experimental device in just the right way to conceal its own existence.* This was the conclusion supported by Hendrik Lorentz. Yes, that's the same guy we named the Lorentz force (defined by Eq. 5.7.1) after. He even performed a mathematical exercise to derive exactly how the aether would have to do this. It was fundamentally the wrong idea, but we'll see later in this chapter that the equations turn out to be correct anyway.

4. *The aether does not exist.* This was highly unappealing at the time because it implies light doesn't need a medium to propagate. It was immediately, but wrongly, discounted as a possible conclusion.

For almost two decades, an argument ensued between supporters of conclusions two and three. The argument wasn't officially settled until Albert


Hendrik Lorentz


Albert Einstein


Hermann Minkowski

Figure 7.1: These people were important in the development of special relativity.

Einstein came along in 1905 (at the age of 26) and published a paper entitled *On the Electrodynamics of Moving Bodies*. In this paper, he presented a rather controversial solution to the problem described in this section that he had been pondering for almost a decade (since the age of about 16). He asked the question that no one else was willing to ask: "What if electrodynamics is completely accurate. but it's classical relativity that needs a bit of reworking?" Needless to say, this solution wasn't well received at the time.

As all hypotheses do, Einstein's included some postulates (i.e. fundamental assumptions). There were only two of these postulates making his idea more elegant than some could be. They involve the concept of **inertial reference frames** (IRFs), which are defined by Newton's first law to be traveling at constant velocity (Note: $\vec{v} = 0$ is constant velocity). Einstein's postulates are:

1. *The laws of physics are the same in all IRFs.* This was nothing new. Having been stated by people like Galileo and Newton, it was over 200 years old in 1905.

2. *The speed of light is constant and the same in all IRFs.* This is the result I mentioned was suggested by Maxwell's equations. Einstein was simply the first to be willing to accept it.

The question that now remains is "If neither the laws of physics nor the speed of light change, then what does change?" The answer is "Almost everything else!" This thought might be a bit difficult to comprehend or accept, but hopefully you'll be able to do both by the end of this chapter.

7.2 Spacetime

When a physics student first learns about special relativity, abstract equations are often thrown at them with little and/or poor explanation. This is a cause for much of the confusion regarding the ideas in this theory. I find it best to build an idea from other ideas a student (or reader) already knows, which is a philosophy I've used in writing this book. We've spent a lot of time focused on coordinate systems and diagrams. This also seems like a good place to start with this.

A major implication of special relativity is that *time* deserves as much attention as *space*. Diagrammatically, that means we'll need to include it in the coordinate system resulting in a four-dimensional **spacetime**. With the new idea of a spacetime comes some new terminology:

- **Spacetime diagram** - A diagram which includes both space and time.
- **Event** - A point in spacetime designated by four coordinates, (ct, x, y, z). Essentially, it's a place and time for some phenomenon.
- **Separation** - The straight line connecting two events in spacetime. The word "distance" is improper with a time component involved.
- **World line** - The path taken by a particle/object in spacetime. The word "trajectory" is improper with a time component involved.

In Figure 7.2, we see two objects initially located at events 1 and 3. At some time Δt later, they are at events 2 and 4, respectively, where they are now closer in space. The line between events 1 and 2 is labeled Δs, which represents the world line of that object. The length of this world line is **spacetime invariant** (i.e. it doesn't change under coordinate transformations).

Line Element

The best tools we have to describe a space are given in Section 6.4. However, we have to be very careful when we incorporate time. First, time is not measured in the same units as space, so a conversion factor of c (the speed of light) appears. Secondly, by observation, we see that time behaves a little

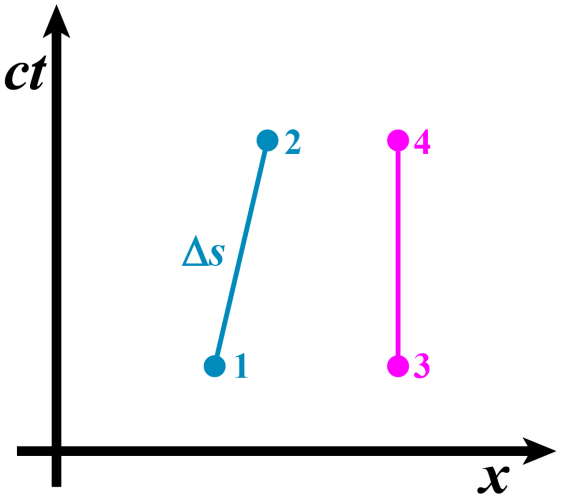


Figure 7.2: This is a spacetime diagram where the horizontal axis, x, represents space (y and z are suppressed for simplicity) and the vertical axis, ct represents time measured in spatial units ($c = 299,792,458$ m/s is like a unit conversion between meters and seconds).

differently than space. It behaves *oppositely* to space, so a negative sign also appears. Keeping all this in mind, the Cartesian line element is now

$$ds^2 = -c^2\,dt^2 + dx^2 + dy^2 + dz^2\,, \tag{7.2.1}$$

which is similar to Eq. 6.4.1. Similar to Eq. 6.4.2, we can write

$$ds^2 = -c^2\,dt^2 + dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2\,, \tag{7.2.2}$$

which is the line element in spherical coordinates. We have simply replaced the spatial terms, with the appropriate dimension-3 line element.

Formulating the mathematics of special relativity in this way was not initially done by Einstein. Einstein's methods involved simple algebra and thought experiments ("Gedankenexperimente" as he called them). He was self-admittedly poor with advanced math. In 1908, Hermann Minkowski generalized Einstein's work with tensor analysis (described in Chapter 6). This is why the space described in this chapter is sometimes called the **Minkowski space**.

Since the labeled world line in Figure 7.2 is straight (true of all world lines in IRFs), we can write it as $(\Delta s)^2 = -c^2\,(\Delta t)^2 + (\Delta x)^2$, which looks a lot like

the Pythagorean theorem by no coincidence. The negative sign on the time component provides some interesting consequences. One consequence is the square of the separation, $(\Delta s)^2$, is not restricted to positive values. We can use this fact to categorize separations in spacetime.

- If $(\Delta s)^2 < 0$, then the two events have a **time-like** separation meaning the time component dominates. All events on world lines showing the motion of massive objects have this kind of separation (considering the large value of c). These world lines are often referred to as time-like world lines.
- If $(\Delta s)^2 = 0$, then the two events have a **light-like** separation because these world lines show the motion of light (and any other massless particle). It is sometimes called a **null** separation because the separation is zero.
- If $(\Delta s)^2 > 0$, then the two events have a **space-like** separation meaning the space component dominates. These two events are considered non-interactive. For an object to travel on a space-like world line, it would require speeds faster than c. For this reason, it is unlikely the motion of anything could be represented by a space-like world line.

From a mathematical standpoint, you could write the time component as an imaginary number since

$$\sqrt{-c^2 (\Delta t)^2} = \sqrt{-1}\, c\Delta t = ic\Delta t.$$

This isn't traditionally done. However, it's mathematically consistent and may be useful under circumstances when you're dealing with the components by themselves rather than in a line element.

Metric Tensor

We can also write something like Eq. 6.4.3 to generalize the line element. The result is

$$ds^2 = g_{\alpha\delta}\, dx^\alpha\, dx^\delta, \tag{7.2.3}$$

where the use of greek indices indicates four dimensions and repeated indices indicates a summation. Remember to distinguish between exponents of 2

and indices! This makes the metric tensor

$$g_{\alpha\delta} \longrightarrow \begin{bmatrix} -c^2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7.2.4}$$

in Cartesian coordinates with an inverse of

$$g^{\alpha\delta} \longrightarrow \begin{bmatrix} -1/c^2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7.2.5}$$

by matrix methods. There is some debate over whether the time component or space components should have the negative sign, but in the end it simply comes down to convention and I've chosen to stick with tradition.

We can transform this to other coordinate systems by replacing the lower-right (spatial) 3×3 with the appropriate dimension-3 metric. For example, in spherical coordinates, we have

$$g_{\alpha\delta} \longrightarrow \begin{bmatrix} -c^2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2\theta \end{bmatrix} \tag{7.2.6}$$

with an inverse metric tensor of

$$g^{\alpha\delta} \longrightarrow \begin{bmatrix} -\frac{1}{c^2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{r^2} & 0 \\ 0 & 0 & 0 & \frac{1}{r^2 \sin^2\theta} \end{bmatrix} \tag{7.2.7}$$

found by matrix methods. Note that we still get $g^{\alpha}_{\delta} = g^{\alpha\mu} g_{\mu\delta} = \delta^{\alpha}_{\delta}$, the same result we got with 3-space in Section 6.4.

Coordinate Rotations

The ultimate value of a spacetime diagram is going to be in how we can use it to look at two different IRFs. Remember from Section 7.1, we're

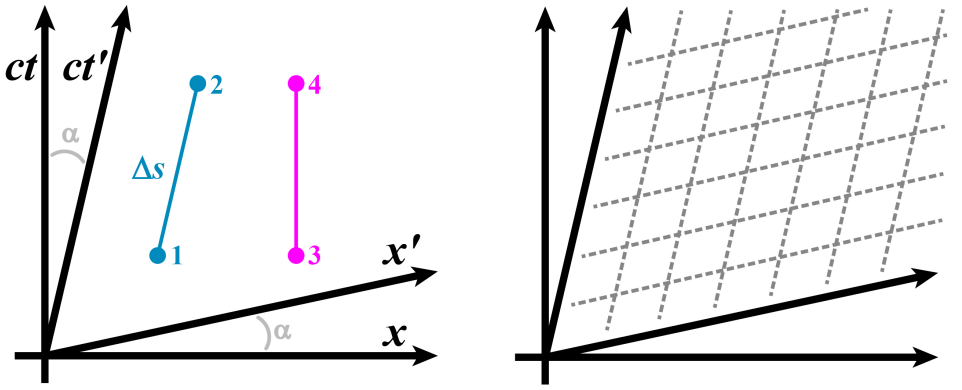


Figure 7.3: In this spacetime diagram, the coordinate systems of both objects are shown as well as both their world lines. Both objects line up with their respective time axis indicating they both consider themselves to be at rest. The diagram on the right shows the grid lines for the primed frame.

trying to explain relative measurements between two perspectives and how this pertains to light. Taking another look at Figure 7.2, we only have one coordinate system shown: the rest frame of the object on the right since it doesn't move in space in that frame (i.e. its world line only has a time component). If we also want to include the rest frame of the object on the left, then we'll need it's time axis to line up with its world line (so it only has a time component in its own frame). That's a coordinate rotation!

However, recall that time and space behave oppositely, so the space axis will have to rotate in the opposite direction. This process is shown in Figure 7.3. The angle, α, shown in the figure is the (circular) angle by which both the axes are rotated between frames on the plane of the paper. Unfortunately, this angle doesn't really tell us much. A rotation in which axes rotate in opposite directions is called a hyperbolic rotation, which involves a hyperbolic angle φ. A hyperbolic angle is really only *analogous* to a circular angle, as we'll see in detail later. You will find we'll discuss α and φ interchangeably with respect to a diagram. This is because α is just a concrete representation of the abstract φ.

In physics, the hyperbolic angle is referred to as the **rapidity**. Just as the $\tan\alpha$ would relate $c\Delta t$ and Δx (and, therefore, v to c) under a normal circular rotation, we have

$$\tanh\varphi = \frac{v}{c} \equiv \beta \tag{7.2.8}$$

for a hyperbolic rotation. We've simply defined β to be v/c (i.e. the fraction of the speed of light). Using a few trigonometric identities, we can solve for cosh and sinh. For instance, we know

$$\tanh^2\varphi = 1 - \operatorname{sech}^2\varphi.$$

With Eq. 7.2.8 and $\operatorname{sech}\varphi = 1/\cosh^2\varphi$, we can say

$$\beta^2 = 1 - \frac{1}{\cosh^2\varphi}$$

$$\cosh\varphi = \frac{1}{\sqrt{1-\beta^2}} \equiv \gamma\,, \tag{7.2.9}$$

where we've simply defined this as a new quantity γ. This γ is referred to as the **gamma factor**. If $\gamma \approx 1$, then the relative motion between the frames is considered classical (i.e. classical physics is within acceptable error). Otherwise, the relative motion between the frames is considered relativistic (i.e. requiring special relativity). Note: If $\beta = 0.14$ (14% of c), then γ is within a percent of 1.

We've found cosh, so what about sinh? Well, there are other trigonometric identities at our disposal. We also know

$$\cosh^2\varphi - \sinh^2\varphi = 1$$

$$\sinh\varphi = \sqrt{\cosh^2\varphi - 1}.$$

With Eq. 7.2.9, we can say

$$\sinh\varphi = \sqrt{\frac{1}{1-\beta^2} - 1}$$

$$\sinh\varphi = \sqrt{\frac{1}{1-\beta^2} - \frac{1-\beta^2}{1-\beta^2}}$$

$$\sinh\varphi = \sqrt{\frac{\beta^2}{1-\beta^2}}$$

$$\sinh\varphi = \frac{\beta}{\sqrt{1-\beta^2}} = \gamma\beta. \tag{7.2.10}$$

Eqs. 7.2.9 and 7.2.10 are very important relationships that will show up repeatedly.

Furthermore, we can get a little more understanding of the diagram out of Eq. 7.2.8. Let's look at our two possible extremes:

- If $v = 0$ ($\beta = 0$), then $\alpha = \varphi = 0$. This makes sense since no relative motion implies no rotation.
- If $v = c$ ($\beta = 1$), then $\alpha = 45°$ and $\varphi = \infty$. This extreme makes it clear that φ is not really an angle in the sense that we typically understand an angle.

In a spacetime diagram, we could say

$$\alpha = \arctan(\tanh\varphi) = \arctan\beta,$$

but this would only be accurate in a diagram like that drawn in Figures 7.3 and 7.4. An axial rotation of $\alpha = 45°$ is just the diagonal exactly between the time and space axes. Events on this diagonal have a light-like separation. Since light is the fastest thing we know of in the universe, we can use this line to define something called a **light cone**.

A light cone points away from an event and encompasses the entire realm of influence of that event on other events in spacetime (and vice versa). Figure 7.4 shows two different light cones for event 1: future (above event 1 in the diagram) and past (below event 1 in the diagram). Event 2 is also on the world line for the object in its future light cone, which means whatever happens there is something the object can come into physical contact with at some point in the future.

Event 3 is on the edge of the future light cone, which means someone at event 3 could *see* the object at event 1, but that's about it. In fact, event 3 would represent an observation of event 1. Event 4 is on the edge of the past light cone, which means the object would receive light from that event (whatever it may be) when it reaches event 1. Also, as time moves forward, the light cone gets larger indicating the light has traveled farther away from where the object was at event 1. Light cones are very useful in discussing the concept of **causality** (i.e. cause and effect).

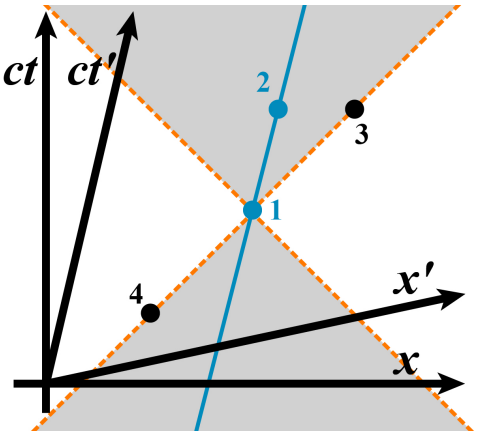


Figure 7.4: In this spacetime diagram, the solid blue line is the world line for an object. The orange dashed lines are world lines for light meaning the shaded triangles represent the past and future light cones of the object at event 1. The cones only appear as triangles due to the suppression of the other spatial coordinates.

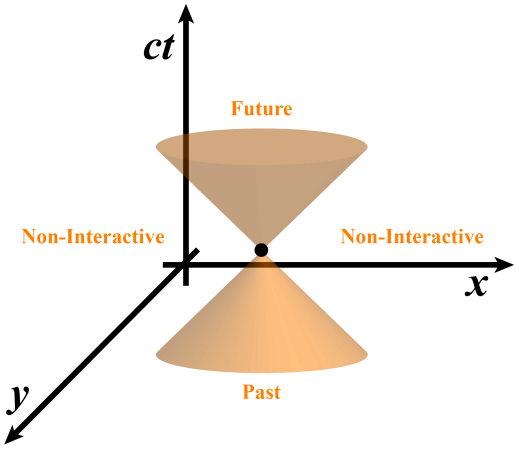


Figure 7.5: This spacetime diagram is very much like Figure 7.4 except only the z-axis is suppressed (rather than both y and z). It is clear in this diagram why we call it a light cone. The event in the center cannot interact with events in the region labeled "Non-Interactive."

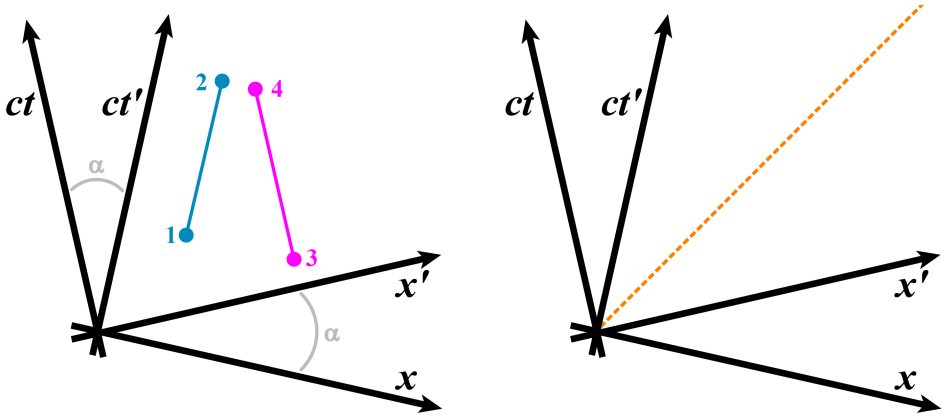


Figure 7.6: In this type of spacetime diagram, neither set of axes looks orthogonal, but it's important to know both sets are orthogonal. The diagram on the left is an exact reproduction of Figure 7.3. On the right, the orange dashed line represents a light-like world line, which is still the diagonal between space and time.

Taking Measurements

It's becoming clear, from diagrams like Figure 7.3, that people in different IRFs will take different measurements (e.g. time and length) of the same phenomenon. This begs the question: "So who's right and who's wrong?" Well, no one is wrong even if different observers don't agree. The concept of absoluteness is something we need to let go in order to move forward in our understanding.

If you have two objects (such as those in Figure 7.3) moving at some relative velocity to one another, then there is no way to determine who is moving and who is not. Object A will consider themselves at rest and say object B is moving (and vice versa). A third observer might say both objects are moving. What we mean is that *all* IRFs are on equal footing. They are all correct about measurements made in their own frame and that's all that matters because of Einstein's first postulate. As long as each observer stays in their own frame, what measurements would be in the other frame is of little consequence.

In Figures 7.3 and 7.4, the unprimed axes are clearly orthogonal to each other. We should take note here that the primed axes are *also* orthogonal to each other even though it's not clear in the diagrams. Sometimes spacetime diagrams are drawn so that neither set of axes looks orthogonal like the one given in Figure 7.6. This helps keep someone working with the topic from

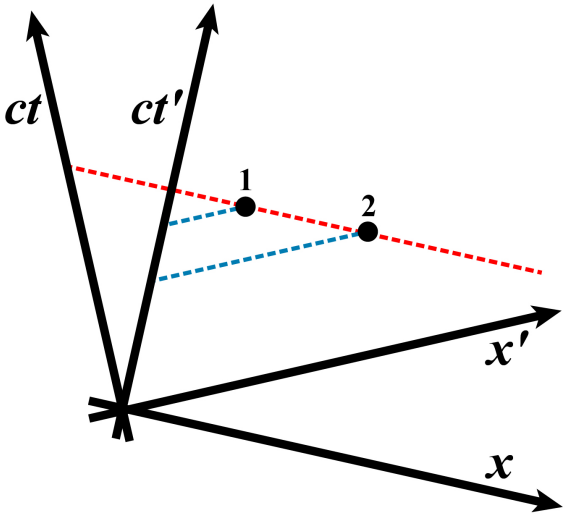


Figure 7.7: In this spacetime diagram, events 1 and 2 are simultaneous in the unprimed frame, but not in the primed frame. Simultaneous events occur in a frame along lines parallel to the spatial axis in that same frame.

giving one IRF preferential treatment.

Another consequence of spacetime relates to **simultaneity**. Just because two events occur at the same time in one IRF, it doesn't mean they occur at the same time in another. This is shown by Figure 7.7 with events 1 and 2. These two events occur at the same time in the unprimed frame as indicated by the downward sloped dashed line. However, in the primed frame, they are separated on the time axis by some $\Delta t'$ (or rather $ic\Delta t'$) as indicated by the upward sloped dashed lines.

Even though, no IRF should ever get preferential treatment, some of them are special for a given a measurement. These are the frames in which the extreme (i.e. maximum or minimum) value is measured. This isn't to say these frames are the *correct* frames, which is sometimes incorrectly implied by calling the measurements **proper quantities**. They're just the frames containing the *extreme* values and some are defined as follows:

- **Proper time**, Δt_p or $\Delta\tau$ - The shortest time.
- **Proper length**, L_p - The longest length.
- **Rest mass**, m_p - The lowest mass.

- **Rest energy**, E_p - The lowest total energy ($E_p = m_p c^2$).

More proper quantities can be defined in terms of these, but it's usually standard to only define the four listed here and let the rest fall as they may.

Example 7.2.1

The difference in time measurements between IRFs is called **time dilation** and we can find it using a spacetime diagram with very little math. For the sake of discussion, let's say a high-speed boat is traveling at night on the ocean at constant v (and, therefore, constant β) in the x-direction. This boat has a blinking light on its bow (safety regulations and all) that blinks at very regular intervals.

Figure 7.8 shows the time dilation in action. Events 1 and 2 represent two consecutive flashes of the boat's bow light. Someone on the boat would be in the primed frame (as this frame is attached to the boat). They measure a spacetime separation of $ic\Delta t'$ between flashes (which is the hypotenuse in the triangle). This is the smallest possible time measurement between these two events, which recall is the proper time ($\Delta t' = \Delta t_p$). You might be inclined to say it's the longest of the three sides of the triangle based on its physical length in the diagram, but don't be fooled! Remember, the time component in the line element is negative.

The green dashed lines are the components of the same world line, but measured in the unprimed frame. The component adjacent to α is measured to be $ic\Delta t$ because it lines up with the ct-axis and the component opposite of α is measured to be Δx because it lines up with the x-axis. It makes sense there would be a Δx in this frame since an observer would see the flashing light move through space.

We can solve this problem one of two ways using the triangle in Figure 7.8. The first instinct might be to use the Pythagorean theorem since the line element looks a lot like it. In that case, we'd get

$$(ic\Delta t')^2 = (ic\Delta t)^2 + (\Delta x)^2$$

$$-c^2 (\Delta t')^2 = -c^2 (\Delta t)^2 + (\Delta x)^2$$

$$c^2 (\Delta t')^2 = c^2 (\Delta t)^2 - (\Delta x)^2 ,$$

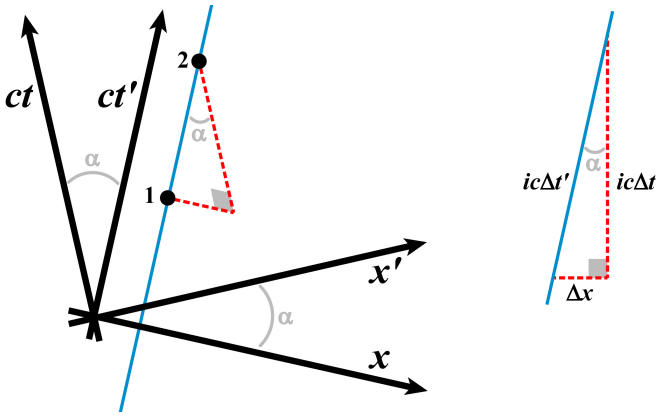


Figure 7.8: In this spacetime diagram, there are two events with time-like separation demonstrating time dilation. The solid blue line is the line element measured as $-c^2\left(\Delta t'\right)^2$. The two red dashed lines represent the components of this same world line measured in the unprimed frame as $-c^2\left(\Delta t\right)^2+\left(\Delta x\right)^2$. The triangle has been straightened-out for clarity.

where we can see $\Delta t' < \Delta t$ due to the subtraction of $\left(\Delta x\right)^2$. This equation also makes sense because we've already said the separation is spacetime invariant. If we divide through by $c^2\left(\Delta t\right)^2$, then the result is

$$\frac{\left(\Delta t'\right)^2}{\left(\Delta t\right)^2} = 1 - \frac{\left(\Delta x\right)^2}{c^2\left(\Delta t\right)^2}$$

$$\left(\frac{\Delta t'}{\Delta t}\right)^2 = 1 - \left(\frac{\Delta x/\Delta t}{c}\right)^2.$$

We know $v = \Delta x/\Delta t$ because the boat has traveled a distance of Δx in a time Δt in the unprimed frame. With this fact and Eq. 7.2.8, we get

$$\left(\frac{\Delta t'}{\Delta t}\right)^2 = 1 - \beta^2$$

$$\frac{\Delta t'}{\Delta t} = \sqrt{1-\beta^2}$$

$$\Delta t = \frac{\Delta t'}{\sqrt{1-\beta^2}}.$$

We can use Eq. 7.2.9 and the definition of proper time to simplify to

$$\boxed{\Delta t = \gamma \Delta t_p} \text{ or } \boxed{\Delta t = \gamma \Delta \tau}, \tag{7.2.11}$$

which is exactly the simple relationship for time dilation.

However, we could have saved a lot of time by using a trigonometric function on the triangle instead. By analogy to circular angles, we get

$$\cosh \varphi = \frac{\text{adjacent}}{\text{hypotenuse}} = \frac{ic\Delta t}{ic\Delta t'} = \frac{\Delta t}{\Delta t'}$$

$$\Delta t = \cosh \varphi \, \Delta t'.$$

Using Eq. 7.2.9 and the definition of proper time, we arrive again at Eq. 7.2.11. This was a much shorter solution, but don't feel bad if you didn't think to do it. Most people aren't familiar enough with hyperbolic trigonometry for it to come to mind. It is something you should put in your arsenal from now on though.

Example 7.2.2

The difference in length measurements between IRFs is called **length contraction** and we can find it using a spacetime diagram with very little math. For the sake of discussion, let's say a high-speed boat is traveling at night on the ocean at constant v (and, therefore, constant β) in the x-direction.

If we're going to measure length, then we need to be clear about what we mean by "length." Measurements of length are very closely related to the concept of simultaneity shown in Figure 7.7. We now define length as the spacetime separation between two particular events. These two events represent the two ends of the object (in this case, the boat). For the measurement to be a length, the two events must occur at the same time in the frame in which you take the measurement.

We've already seen that simultaneous events in one IRF are not simultaneous in any another IRF, so the set of events measuring length in one frame will not be the same set of events measuring length in the other. Figure 7.9 shows the length contraction in action. Length in the primed frame is

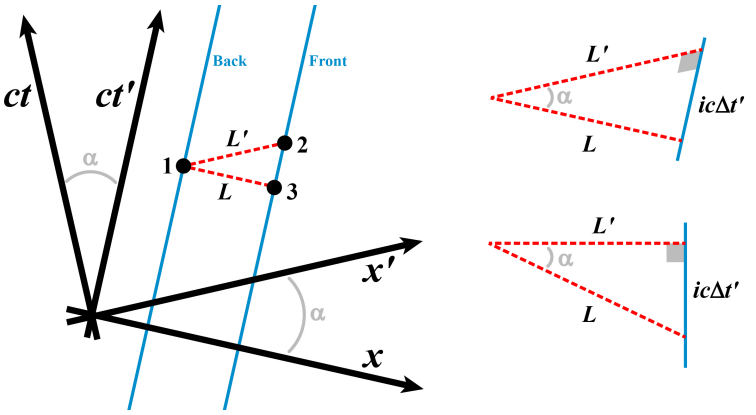


Figure 7.9: In this spacetime diagram, there are two world lines corresponding the front and back of an object. Between them, there are two measurements of length corresponding to the two different frames. Both must connect the two world lines with events which occur at the same time in the frame of measurement. The top triangle is an enlarged version of the one in the diagram and the bottom triangle is just a straightened-out version for clarity.

measured between events 1 and 2, where as length in the unprimed frame is measured between events 1 and 3. The sets are only allowed to have one event in common.

We can perform a little hyperbolic trigonometry on the triangle in Figure 7.9 just as we did with Example 7.2.1. This results in

$$\cosh\varphi = \frac{\text{adjacent}}{\text{hypotenuse}} = \frac{L'}{L}$$

$$L = \frac{L'}{\cosh\varphi}$$

Using Eq. 7.2.9 and the definition of proper length, we arrive

$$\boxed{L = \frac{L_p}{\gamma}}. \tag{7.2.12}$$

You might be inclined to say L the longest of the three sides of the triangle based on its physical length in the diagram, so you'd think it would be the longer length measurement. Don't be fooled! Remember, the square of time component is negative, so the Pythagorean theorem says

$$(L)^2 = (ic\Delta t')^2 + (L')^2$$

$$(L)^2 = -c^2 (\Delta t')^2 + (L')^2 ,$$

where we can clearly see $L' > L$ due to the subtraction of $c^2 (\Delta t')^2$. Furthermore, it's important to know that both these lengths are measured in the direction of motion. There is no length contraction along the other orthogonal directions (i.e. the y and z directions).

7.3 Lorentz Transformations

In Section 7.1, we mentioned Hendrik Lorentz and his idea that the luminiferous aether somehow contracted experimental devices to conceal its own existence. This was, and still is, a preposterous idea. However, the equations he derived for the process turn out to be exactly the equations Einstein derived (with more sound fundamental concepts). These equations are actually a coordinate transformation from one IRF to another. Rather than give you traditional derivation in this book, I have opted to derive them using the method of spacetime diagrams described in Section 7.2.

We've mentioned that moving to a set of coordinates in another IRF is represented by a hyperbolic rotation in a spacetime diagram. Let's start this discussion from the standpoint of a normal circular rotation in 3-space. We can use a rotation matrix to rotate spatial axes (as done in Example 6.5.1). For example,

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

rotates the coordinate system couterclockwise about the z-axis. The value of 1 in the matrix shows that the z-component does not change under a rotation about the z-axis (i.e. only the x and y components change).

Under the hyperbolic rotation in spacetime, only the space axis along the direction of motion (we'll call it x) and the time axis rotate, where the other two space axes do not. In matrix form, we'd say

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cosh\varphi & -\sinh\varphi & 0 & 0 \\ -\sinh\varphi & \cosh\varphi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix} ,$$

which transforms coordinates in spacetime (hence four components rather than three). This corresponds to a *counterclockwise* rotation of the x-axis (and a clockwise rotation of the ct-axis). In other words, the primed frame is moving in the positive x-direction according to the unprimed frame (the frame on which the transformation takes place). A transformation in the other direction (i.e. the other frame is perceived to move in the negative x-direction) will require the inverse matrix or, put more simply: replace $-\sinh\varphi$ with $\sinh\varphi$ (i.e. *clockwise* for the x-axis).

We can get away from the rapidity notation by taking advantage of Eqs. 7.2.9 and 7.2.10. Therefore,

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix}, \tag{7.3.1}$$

which looks a lot simpler and is more oriented toward measurable values (noting again that $-\beta$ is replaced by β for the inverse transformation). If you prefer transformation equations over matrices, then we can just perform a quick matrix multiplication. Eq. 7.3.1 becomes

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \gamma ct - \gamma\beta x \\ -\gamma\beta ct + \gamma x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \gamma\left(ct - \beta x\right) \\ \gamma\left(-\beta ct + x\right) \\ y \\ z \end{bmatrix}.$$

We can divide the first line by c and use Eq. 7.2.8 to get

$$\begin{Bmatrix} t' &=& \gamma\left(t - vx/c^2\right) \\ x' &=& \gamma\left(-vt + x\right) \\ y' &=& y \\ z' &=& z \end{Bmatrix}, \tag{7.3.2}$$

which is the familiar form from an introductory textbook. However, I would highly recommend the matrix or index method as they drastically simplify the math.

Example 7.3.1

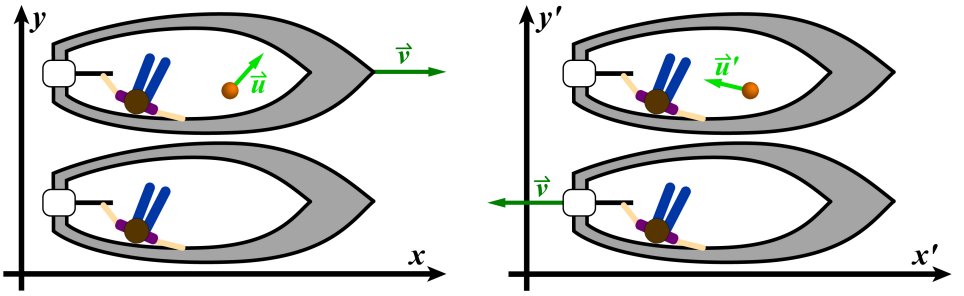


Figure 7.10: A ball is thrown in the top boat. In the unprimed frame (attached to the bottom boat), the top boat is moving in the positive x-direction at v and the velocity of the ball is measured to be $\vec{u}$. In the primed frame (attached to the top boat), the bottom boat is moving in the negative x-direction at v and the velocity of the ball is measured to be $\vec{u}'$.

You're on the ocean on a boat at rest (relative to the water) when you see a high-speed boat zip passed you. It is traveling at constant v (and, therefore, constant β) in the x-direction. At that exact moment, the driver of that other boat throws a ball in a random direction with a velocity you measure to be $\vec{u}$ (pun intended). What velocity would the driver of the other boat measure for the ball?

- We can do this component-wise, so let's start with the x-direction (the boat's direction of motion). The definition of velocity in this direction is

$$u'_x = \frac{dx'}{dt'}.$$

  We can apply Eq. 7.3.2 to both the numerator and denominator (as they both change between IRFs). The result is

$$u'_x = \frac{\gamma\,(-v\,dt + dx)}{\gamma\,(dt - v\,dx/c^2)}.$$

  Dividing the numerator and denominator by $\gamma\,dt$ gives us

$$u'_x = \frac{-v + dx/dt}{1 - v\,dx/\,(c^2\,dt)} = \frac{dx/dt - v}{1 - (dx/dt)\,v/c^2}.$$

  We know $u_x = dx/dt$, so

$$u'_x = \frac{u_x - v}{1 - u_x v/c^2}.$$

- Performing this same process for the y-direction, we get

$$u'_y = \frac{dy'}{dt'} = \frac{dy}{\gamma\,(dt - v\,dx/c^2)}$$

$$u'_y = \frac{dy/dt}{\gamma\,[1 - v\,dx/\,(c^2\,dt)]} = \frac{dy/dt}{\gamma\,[1 - (dx/dt)\,v/c^2]}.$$

$$u'_y = \frac{u_y/\gamma}{1 - u_x v/c^2}.$$

  We get something very similar for the z-direction.

In summary,

$$\boxed{u'_x = \frac{u_x - v}{1 - u_x v/c^2}} \tag{7.3.3a}$$

$$\boxed{u'_y = \frac{u_y/\gamma}{1 - u_x v/c^2}} \tag{7.3.3b}$$

$$\boxed{u'_z = \frac{u_z/\gamma}{1 - u_x v/c^2}} \tag{7.3.3c}$$

where x is the direction of motion of the primed IRF relative to the unprimed IRF. This is called **coordinate velocity** since the derivative is taken with respect to the coordinate time, t. According to the observer on the other boat, they are at rest and you're moving in the negative x-direction as shown in Figure 7.10. That means you can obtain the reverse transformations (i.e. $\vec{u}' \to \vec{u}$) by replacing v with $-v$. Note that $\vec{u}$ and $\vec{u}'$ need not have the same magnitude nor the same direction.

Example 7.3.2

You're on the ocean on a boat at rest (relative to the water) when you see a high-speed boat zip passed you. It is traveling at constant v (and, therefore, constant β) in the x-direction. At that exact moment, the driver of that other boat throws a ball in a random direction with an acceleration you measure to be $\vec{a}$. What acceleration would the driver of the other boat measure for the ball?

- We can do this component-wise, so let's start with the x-direction (the boat's direction of motion). The definition of velocity in this direction is

$$a'_x = \frac{du'_x}{dt'} = \frac{du'_x/dt}{dt'/dt}.$$

We can apply Eq. 7.3.2 to both the denominator and Eq. 7.3.3a to numerator. The result is

$$a'_x = \frac{\dfrac{d}{dt}\left(\dfrac{u_x - v}{1 - u_x v/c^2}\right)}{\gamma \dfrac{d}{dt}\left(t - \dfrac{vx}{c^2}\right)}.$$

where γ and v are both constant. Using the derivative quotient rule (defined by Eq. 3.1.6) on the numerator and distributing the denominator gives us

$$a'_x = \frac{\dfrac{du_x}{dt}\left(1 - \dfrac{u_x v}{c^2}\right) - (u_x - v)\dfrac{d}{dt}\left(1 - \dfrac{u_x v}{c^2}\right)}{\left(1 - \dfrac{u_x v}{c^2}\right)^2} \left[\frac{1}{\gamma\left(\dfrac{dt}{dt} - \dfrac{v}{c^2}\dfrac{dx}{dt}\right)}\right]$$

$$a'_x = \frac{\dfrac{du_x}{dt}\left(1 - \dfrac{u_x v}{c^2}\right) - (u_x - v)\dfrac{-v}{c^2}\dfrac{du_x}{dt}}{\left(1 - \dfrac{u_x v}{c^2}\right)^2} \left[\frac{1}{\gamma\left(1 - \dfrac{v}{c^2}\dfrac{dx}{dt}\right)}\right].$$

We know $u_x = dx/dt$ and $a_x = du_x/dt$, so

$$a'_x = \frac{a_x\left(1 - \dfrac{u_x v}{c^2}\right) - (u_x - v)\dfrac{-v}{c^2}a_x}{\left(1 - \dfrac{u_x v}{c^2}\right)^2} \left[\frac{1}{\gamma\left(1 - \dfrac{u_x v}{c^2}\right)}\right]$$

$$a'_x = \frac{\left(1 - \frac{u_x v}{c^2}\right) - (u_x - v)\frac{-v}{c^2}}{\gamma\left(1 - \frac{u_x v}{c^2}\right)^3} a_x$$

$$a'_x = \frac{1 - \frac{u_x v}{c^2} + \frac{u_x v}{c^2} - \frac{v^2}{c^2}}{\gamma\left(1 - \frac{u_x v}{c^2}\right)^3} a_x = \frac{1 - \frac{v^2}{c^2}}{\gamma\left(1 - \frac{u_x v}{c^2}\right)^3} a_x$$

and, by Eq. 7.2.9,

$$a'_x = \frac{a_x}{\gamma^3\left(1 - u_x v/c^2\right)^3}$$

- Performing this same process for the y-direction, we get

$$a'_y = \frac{du'_y}{dt'} = \frac{du'_y/dt}{dt'/dt} = \frac{\frac{d}{dt}\left(\frac{u_y/\gamma}{1 - u_x v/c^2}\right)}{\gamma\frac{d}{dt}\left(t - \frac{vx}{c^2}\right)}$$

$$a'_y = \frac{\frac{du_y}{dt}\left(1 - \frac{u_x v}{c^2}\right) - u_y\frac{d}{dt}\left(1 - \frac{u_x v}{c^2}\right)}{\left(1 - \frac{u_x v}{c^2}\right)^2}\left[\frac{1}{\gamma^2\left(\frac{dt}{dt} - \frac{v}{c^2}\frac{dx}{dt}\right)}\right]$$

$$a'_y = \frac{\frac{du_y}{dt}\left(1 - \frac{u_x v}{c^2}\right) - u_y\frac{-v}{c^2}\frac{du_x}{dt}}{\left(1 - \frac{u_x v}{c^2}\right)^2}\left[\frac{1}{\gamma^2\left(1 - \frac{v}{c^2}\frac{dx}{dt}\right)}\right]$$

$$a'_y = \frac{a_y\left(1 - \frac{u_x v}{c^2}\right) + \frac{u_y v}{c^2}a_x}{\left(1 - \frac{u_x v}{c^2}\right)^2}\left[\frac{1}{\gamma^2\left(1 - \frac{u_x v}{c^2}\right)}\right]$$

$$a'_y = \frac{\left(1 - u_x v/c^2\right)a_y + \left(u_y v/c^2\right)a_x}{\gamma^2\left(1 - u_x v/c^2\right)^3}.$$

We get something very similar for the z-direction.

In summary,

$$\boxed{a'_x = \frac{a_x}{\gamma^3 \left(1 - u_x v/c^2\right)^3}} \tag{7.3.4a}$$

$$\boxed{a'_y = \frac{\left(1 - u_x v/c^2\right) a_y + \left(u_y v/c^2\right) a_x}{\gamma^2 \left(1 - u_x v/c^2\right)^3}} \tag{7.3.4b}$$

$$\boxed{a'_z = \frac{\left(1 - u_x v/c^2\right) a_z + \left(u_z v/c^2\right) a_x}{\gamma^2 \left(1 - u_x v/c^2\right)^3}} \tag{7.3.4c}$$

where x is the direction of motion of the primed IRF relative to the unprimed IRF. This is called **coordinate acceleration** since the derivative is taken with respect to the coordinate time, t. You can see that Eqs. 7.3.4b and 7.3.4c are also dependent on a_x, which makes these transformations very complicated. According to the observer on the other boat, they are at rest and you're moving in the negative x-direction as shown in Figure 7.10. That means you can obtain the reverse transformations (i.e. $\vec{a}' \to \vec{a}$) by replacing v with $-v$. Note that $\vec{a}$ and $\vec{a}'$ need not have the same magnitude nor the same direction.

Transformation Matrix

The 4×4 matrix in Eq. 7.3.1 is called the Lorentz transformation matrix and is given by

$$\Lambda^\alpha_\delta \longrightarrow \begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{7.3.5}$$

where we've used a capital lambda to represent it (noting again that $-\beta$ is replaced by β for the inverse transformation). We can actually write Eq. 7.3.1 in index notation using this Λ matrix and changing the contravariant coordinates from (ct, x, y, z) to (x^0, x^1, x^2, x^3). Under this notation, it becomes

$$x'^\alpha = \Lambda^\alpha_\delta x^\delta \,. \tag{7.3.6}$$

Notice, we made the definition $x^0 \equiv ct$ that merges the quantity c into the time component. We're now measuring time in spatial units (e.g. meters). This changes the look of our line element in Cartesian coordinates to

$$g_{\alpha\delta}\, dx^\alpha\, dx^\delta = -dx^0 dx^0 + dx^1 dx^1 + dx^2 dx^2 + dx^3 dx^3 \tag{7.3.7}$$

and the spacetime metric tensor to

$$g_{\alpha\delta} \longrightarrow \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{7.3.8}$$

where we would still replace the lower right 3×3 components with the appropriate 3-space metric. This definition of the time component comes with its conveniences. First, the metric is not only simpler, but it reflects clearly our choice of sign convention $(-,+,+,+)$. Secondly, we don't have to worry about factors of c appearing in equations and when we raise or lower indices. The only downside is we must think about time differently, which isn't an unreasonable expectation given that spacetime puts space and time on equal footing. If you think about it, we're already accustomed to the reverse, measuring space with time units: "The store is 15 minutes away."

Inconvenient Coordinates

We can also extend Eq. 7.3.5 a bit. So far, we've been assuming that the relative motion between the two IRFs is solely in the x-direction (i.e. $v_y = v_z = 0$). This wasn't an unrealistic assumption, mind you, since constant velocity (direction included) implies linear motion. Unfortunately, some systems are complex enough that another phenomenon may dictate the location and orientation of the coordinate system. If that's the case, we'll need to generalize Eq. 7.3.5 to a relative velocity with three non-zero components.

We can see from Eq. 7.3.2 that the directions orthogonal to the direction of motion are unaffected by the transformation. This should still be true if we generalize since the orientation of the coordinate system should not affect physical results. With that in mind, let's consider the dimension-3 position vector of an event. We can split this into two components: one parallel to $\vec{v}$ and one perpendicular to $\vec{v}$ such that

$$\vec{r} = \vec{r}_\parallel + \vec{r}_\perp. \tag{7.3.9}$$

Eq. 7.3.2 now becomes

$$\left.\begin{array}{rcl} t' &=& \gamma\left(t - \vec{v}\bullet\vec{r}_{\parallel}/c^2\right) \\ \vec{r}'_{\parallel} &=& \gamma\left(-\vec{v}t + \vec{r}_{\parallel}\right) \\ \vec{r}'_{\perp} &=& \vec{r}_{\perp} \end{array}\right\},$$

where we've replaced v and x with $\vec{v}$ and $\vec{r}_{\parallel}$, respectively. Since $\vec{v}\bullet\vec{r}_{\perp} = 0$ by definition of $\vec{r}_{\perp}$, then we can use Eq. 7.3.9 to get

$$\vec{v}\bullet\vec{r}_{\parallel} = \vec{v}\bullet\left(\vec{r} - \vec{r}_{\perp}\right) = \vec{v}\bullet\vec{r}.$$

For consistent units, we can also multiply the top equation by c arriving at

$$ct' = \gamma\left(ct - \vec{v}\bullet\vec{r}/c\right).$$

We also need a substitution for $\vec{r}_{\parallel}$ and $\vec{r}_{\perp}$. The parallel component can be written as a projection onto $\vec{v}$ by

$$\vec{r}_{\parallel} = \left(\hat{v}\bullet\vec{r}\right)\hat{v} = \left(\frac{\vec{v}}{v}\bullet\vec{r}\right)\frac{\vec{v}}{v} = \frac{\left(\vec{v}\bullet\vec{r}\right)\vec{v}}{v^2}$$

where $\hat{v}$ is the unit vector in the direction of motion and we've used something like Eq. 5.2.4 to get rid of the unit vectors.

It's going to be simpler in the long run to use β rather than v, so we'll define $\vec{\beta} = \vec{v}/c$. We now have

$$\vec{r}_{\parallel} = \frac{\left(\vec{\beta}\bullet\vec{r}\right)\vec{\beta}}{\beta^2}$$

and the perpendicular component follows from Eq. 7.3.9 as

$$\vec{r}_{\perp} = \vec{r} - \vec{r}_{\parallel} = \vec{r} - \frac{\left(\vec{\beta}\bullet\vec{r}\right)\vec{\beta}}{\beta^2}.$$

Therefore, the transformation equations are

$$\left.\begin{array}{rcl} ct' &=& \gamma\left(ct - \vec{\beta}\bullet\vec{r}\right) \\ \vec{r}'_{\parallel} &=& \gamma\left[-\vec{\beta}ct + \dfrac{\left(\vec{\beta}\bullet\vec{r}\right)\vec{\beta}}{\beta^2}\right] \\ \vec{r}'_{\perp} &=& \vec{r} - \dfrac{\left(\vec{\beta}\bullet\vec{r}\right)\vec{\beta}}{\beta^2} \end{array}\right\}.$$

Lastly, we can merge the last two equations by using Eq. 7.3.9 in the primed system. This gives us

$$\left\{\begin{array}{rcl} ct' &=& \gamma\left(ct - \vec{\beta}\bullet\vec{r}\right) \\ \vec{r}' &=& -\gamma\vec{\beta}ct + \vec{r} + (\gamma-1)\dfrac{\left(\vec{\beta}\bullet\vec{r}\right)\vec{\beta}}{\beta^2} \end{array}\right\},$$

or even better yet

$$\left\{\begin{array}{rcl} ct' &=& \gamma\left(ct - \beta_x x - \beta_y y - \beta_z z\right) \\ x' &=& -\gamma\beta_x ct + x + (\gamma-1)\dfrac{\left(\beta_x x + \beta_y y + \beta_z z\right)\beta_x}{\beta^2} \\ y' &=& -\gamma\beta_y ct + y + (\gamma-1)\dfrac{\left(\beta_x x + \beta_y y + \beta_z z\right)\beta_y}{\beta^2} \\ z' &=& -\gamma\beta_z ct + z + (\gamma-1)\dfrac{\left(\beta_x x + \beta_y y + \beta_z z\right)\beta_z}{\beta^2} \end{array}\right\}$$

$$\left\{\begin{array}{rcl} ct' &=& \gamma ct - \gamma\beta_x x - \gamma\beta_y y - \gamma\beta_z z \\ x' &=& -\gamma\beta_x ct + \left[1 + (\gamma-1)\dfrac{\beta_x^2}{\beta^2}\right]x + (\gamma-1)\dfrac{\beta_x\beta_y}{\beta^2}y + (\gamma-1)\dfrac{\beta_x\beta_z}{\beta^2}z \\ y' &=& -\gamma\beta_y ct + (\gamma-1)\dfrac{\beta_y\beta_x}{\beta^2}x + \left[1 + (\gamma-1)\dfrac{\beta_y^2}{\beta^2}\right]y + (\gamma-1)\dfrac{\beta_y\beta_z}{\beta^2}z \\ z' &=& -\gamma\beta_z ct + (\gamma-1)\dfrac{\beta_z\beta_x}{\beta^2}x + (\gamma-1)\dfrac{\beta_z\beta_y}{\beta^2}y + \left[1 + (\gamma-1)\dfrac{\beta_z^2}{\beta^2}\right]z \end{array}\right\},$$

where we've used Eq. 2.2.2 to expand the dot products. This makes the general transformation matrix

$$\Lambda^\alpha_\delta \rightarrow \begin{bmatrix} \gamma & -\gamma\beta_x & -\gamma\beta_y & -\gamma\beta_z \\ -\gamma\beta_x & 1 + (\gamma-1)\dfrac{\beta_x^2}{\beta^2} & (\gamma-1)\dfrac{\beta_x\beta_y}{\beta^2} & (\gamma-1)\dfrac{\beta_x\beta_z}{\beta^2}z \\ -\gamma\beta_y & (\gamma-1)\dfrac{\beta_y\beta_x}{\beta^2} & 1 + (\gamma-1)\dfrac{\beta_y^2}{\beta^2} & (\gamma-1)\dfrac{\beta_y\beta_z}{\beta^2} \\ -\gamma\beta_z & (\gamma-1)\dfrac{\beta_z\beta_x}{\beta^2} & (\gamma-1)\dfrac{\beta_z\beta_y}{\beta^2} & 1 + (\gamma-1)\dfrac{\beta_z^2}{\beta^2} \end{bmatrix} \quad (7.3.10)$$

which still obeys Eq. 7.3.6.

It's important to point out here that Eq. 7.3.10 only applies to Cartesian coordinates and only when the two system (primed and unprimed) have parallel unit vectors. If either of these two conditions isn't met, then the transformation matrix is *far* more complicated. Furthermore, the coordinate velocities and accelerations we found to be Eq. Sets 7.3.3 and 7.3.4, respectively, also get proportionally more complicated with the transformation matrix. Luckily, special relativity dictates constant motion between frames to which a Cartesian coordinate system lends itself quite nicely.

7.4 Relativistic Dynamics

We've been discussing special relativity as though it's an entire set of mechanics. Ideally, we'd like to be able to easily transform all vectors (e.g. velocity, acceleration, momentum, or force) from one frame to another. We can do this as long as we're careful.

In Section 6.6, we discussed transformations of all kinds of tensors (vectors included). Eqs. 6.6.2 and/or 6.6.3 governed how dimension-3 tensors transformed from one set of coordinates to another. However, pseudotensors were a completely different story. If we plan to use Eq. 7.3.6 to transform **4-vectors** (i.e. dimension-4 vectors) in spacetime, then we'd better make sure they're real 4-vectors rather than 4-pseudovectors. This would take the form

$$T'^{\alpha} = \Lambda^{\alpha}_{\delta} T^{\delta} \tag{7.4.1}$$

such that T^{δ} is an arbitrary 4-vector.

A simple example of a real vector in spacetime is the displacement 4-vector (or 4-displacement). It represents the separation between two events in spacetime and its contravariant form is given by

$$\Delta x^{\alpha} \longrightarrow \begin{bmatrix} \Delta x^0 \\ \Delta x^1 \\ \Delta x^2 \\ \Delta x^3 \end{bmatrix} = \begin{bmatrix} c\Delta t \\ \Delta x \\ \Delta y \\ \Delta z \end{bmatrix},$$

where $x^0 = ct$. The 4-displacement is physically more important than the 4-position because one event in spacetime doesn't really mean much. For a

phenomenon to mean anything to us, we must at least observe it, which in itself is a second event. Furthermore, a zero 4-displacement means something physical: the events happened at the same time and place. The same can not be said for the 4-position since the origin can be placed anywhere without affecting the real physical world.

If we take the scalar product (a generalized dot product) of the 4-displacement with itself, then it will be

$$\Delta x_\alpha \Delta x^\alpha = g_{\delta\alpha} \Delta x^\delta \Delta x^\alpha,$$

where we have used the metric tensor to raise the index on the first vector in the product. This looks a lot like the line element in Eq. 7.2.3. Using Eq. 7.3.8, we get

$$\Delta x_\alpha \Delta x^\alpha = -\Delta x^0 \Delta x^0 + \Delta x^1 \Delta x^1 + \Delta x^2 \Delta x^2 + \Delta x^3 \Delta x^3$$

$$\boxed{\Delta x_\alpha \Delta x^\alpha = -c^2 (\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2}, \tag{7.4.2}$$

which is almost exactly the line element in Cartesian coordinates. The only difference is the Δ instead of the differential, but you could just as easily do this for an infinitesimally small displacement: $dx_\alpha dx^\alpha$. This makes the scalar product of the 4-displacement with itself a spacetime invariant (i.e. $\Delta x_\alpha \Delta x^\alpha = \Delta x'_\delta \Delta x'^\delta$), which is something we mentioned in Section 7.2.

It is important to note that the covariant 4-displacement is given by

$$\Delta x_\alpha = g_{\delta\alpha} \Delta x^\delta \longrightarrow \begin{bmatrix} -\Delta x^0 \\ \Delta x^1 \\ \Delta x^2 \\ \Delta x^3 \end{bmatrix} = \begin{bmatrix} -c\Delta t \\ \Delta x \\ \Delta y \\ \Delta z \end{bmatrix},$$

where the only difference between this and the contravariant form is the negative on the time component. This mathematical phenomenon is true of all 4-vectors due to the metric tensor in Eq. 7.3.8. Sometimes it is written shorthand as $(-c\Delta t, \Delta\vec{r})$ where $\Delta\vec{r}$ is the 3-displacement. In general, this shorthand is essentially $(\text{time}, \overrightarrow{\text{space}})$.

Unlike classical physics, time derivatives of 4-vectors are not necessarily also 4-vectors. Time is measured differently in different IRFs, which poses issues. You also can't take a 3-vector, just tack on a fourth component,

and call it a 4-vector. For example, the 3-velocity extended into dimension-4 would have a time component of $dx^0/dt = c$, but it's a 4-pseudovector. This is something made clear by the look of the transformations in Example 7.3.1. To make a 4-vector by taking a time derivative, we need to use a time measurement all frames can agree on. Traditionally, we go with the proper time, $\Delta\tau$.

Four-Velocity

If we're talking time derivatives, then it makes sense to start with velocity. The dimension-4 velocity vector of an object is defined as

$$u^\delta = \frac{dx^\delta}{d\tau}, \tag{7.4.3}$$

the first derivative of 4-position with respect to proper time. It is commonly called the **4-velocity** and we can make it look a little more familiar. If we use the chain rule (defined by Eq. 3.1.2) and time dilation (defined by Eq. 7.2.11), then

$$u^\delta = \frac{dx^\delta}{dt}\frac{dt}{d\tau} = \gamma\frac{dx^\delta}{dt},$$

where

$$\gamma = \frac{1}{\sqrt{1 - u^2/c^2}} \tag{7.4.4}$$

and $\vec{u}$ is the relative velocity between the object and the frame in which its velocity is measured. The velocity $\vec{u}$ is *not* the same as the relative velocity $\vec{v}$ between two observers in two different IRFs. The object itself would represent a third frame (not necessarily an IRF) independent from the other two where it measures its own proper time. It's components in Cartesian coordinates can be shown in matrix notation as

$$u^\delta \longrightarrow \begin{bmatrix} \gamma\, dx^0/dt \\ \gamma\, dx^1/dt \\ \gamma\, dx^2/dt \\ \gamma\, dx^3/dt \end{bmatrix} = \begin{bmatrix} \gamma c\, dt/dt \\ \gamma\, dx/dt \\ \gamma\, dy/dt \\ \gamma\, dz/dt \end{bmatrix} = \begin{bmatrix} \gamma c \\ \gamma u_x \\ \gamma u_y \\ \gamma u_z \end{bmatrix}, \tag{7.4.5}$$

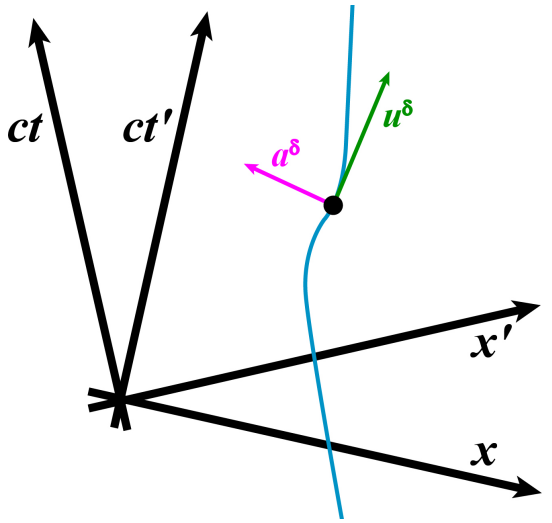


Figure 7.11: In this spacetime diagram, the world line for an object is shown. Its 4-velocity, u^δ, is indictated by a green arrow and its 4-acceleration, a^δ, is shown with a purple arrow.

where $\vec{u} = (u_x, u_y, u_z)$ is the coordinate 3-velocity described in Example 7.3.1. It can also be written in shorthand as $(\gamma c, \gamma\vec{u})$.

The 4-velocity can be looked at another way. As a proper time derivative of 4-position, it represents the tangent vector to the object's world line (see Figure 7.11). This world line does *not* have to be straight because its own frame doesn't have to be an IRF. Therefore, Eq. 7.4.4 doesn't have to be constant.

An interesting quality of the 4-velocity can be show from the scalar product with itself. It is given by

$$u_\delta u^\delta = u_0 u^0 + u_1 u^1 + u_2 u^2 + u_3 u^3 = u_0 u^0 + \gamma^2 \vec{u} \bullet \vec{u},$$

where we've written the spatial product as the familiar dot product (think shorthand 4-vector notation). We also know $u_0 = g_{0\mu} u^\mu = g_{00} u^0 = -u^0$ because the metric tensor is diagonal. Now the scalar product is

$$u_\delta u^\delta = (-\gamma c)(\gamma c) + \gamma^2 \vec{u} \bullet \vec{u}$$

$$u_\delta u^\delta = -\gamma^2 c^2 + \gamma^2 u^2 = -\gamma^2 c^2 \left(1 - \frac{u^2}{c^2}\right)$$

$$u_\delta u^\delta = -c^2, \tag{7.4.6}$$

which is constant and true for all 4-velocities in all time-like frames. We could have also said

$$u_\delta u^\delta = g_{\delta\mu} u^\mu u^\delta = g_{\delta\mu} \frac{dx^\mu}{d\tau} \frac{dx^\delta}{d\tau} = \frac{g_{\delta\mu} dx^\mu dx^\delta}{d\tau d\tau},$$

by Eq. 7.4.3. The numerator is just the general definition of the line element, so

$$u_\delta u^\delta = \frac{-c^2 d\tau^2}{d\tau d\tau} = -c^2,$$

where we've assumed $ds^2 = -c^2 d\tau^2$ in the rest frame of the object (see Example 7.2.1). This is exactly the same result as before.

You could argue the magnitude of the 4-velocity for all particles is $\sqrt{u_\delta u^\delta} = ic$ and it's only the components that IRFs measure differently. In the rest frame of the object, the contravariant 4-velocity is $(c, 0)$, which is a fact we can use to derive the generalized 4-velocity another way. If we use a Lorentz transformation from the rest frame of the object into an arbitrary IRF, the result is

$$u^\delta \longrightarrow \begin{bmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \gamma c \\ \gamma\beta c \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \gamma c \\ \gamma u \\ 0 \\ 0 \end{bmatrix}, \tag{7.4.7}$$

where γ is given by Eq. 7.4.4 and we're assuming β is positive due to the direction of transformation. Note: The result would have been exactly Eq. 7.4.5 had we used Eq. 7.3.10 as the transformation matrix instead. You'll find this method is a very useful short-cut to have in your mathematical toolbox.

We can also use the transformation of 4-velocity to write out a transformation for the coordinate velocity 4-pseudovector. Between two arbitrary IRFs, the transformation for the 4-velocity is

$$u'^\delta = \Lambda^\delta_\mu u^\delta \tag{7.4.8}$$

$$\begin{bmatrix} \gamma' c \\ \gamma' u'_x \\ \gamma' u'_y \\ \gamma' u'_z \end{bmatrix} = \begin{bmatrix} \gamma_T & -\gamma_T\beta_T & 0 & 0 \\ -\gamma_T\beta_T & \gamma_T & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma c \\ \gamma u_x \\ \gamma u_y \\ \gamma u_z \end{bmatrix}$$

where the T subscript stands for "transformation." There are three different gammas: γ is between the unprimed frame and the objects rest frame, γ' is between the primed frame and the objects rest frame, and γ_T is between the primed and unprimed frames. If we move around the γ and γ', then we have

$$\begin{bmatrix} c \\ u'_x \\ u'_y \\ u'_z \end{bmatrix} = \left(\frac{\gamma}{\gamma'}\right) \begin{bmatrix} \gamma_T & -\gamma_T\beta_T & 0 & 0 \\ -\gamma_T\beta_T & \gamma_T & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c \\ u_x \\ u_y \\ u_z \end{bmatrix}$$

and now the column matrices are just the coordinate velocity 4-pseudovectors in primed and unprimed frames. Now we can write

$$\frac{dx'^{\delta}}{dt'} = \left(\frac{\gamma}{\gamma'}\right) \Lambda^{\delta}_{\mu} \frac{dx^{\mu}}{dt}, \tag{7.4.9}$$

which is very reminiscent of a pseudovector transformation given that it simply gains an extra scalar factor.

Four-Acceleration

If an object is not only moving but accelerating, then we'll also need to a second derivative of its 4-position. We call this the **4-acceleration** and it is defined by

$$a^{\delta} = \frac{d}{d\tau}\left(\frac{dx^{\delta}}{d\tau}\right) = \frac{du^{\delta}}{d\tau}. \tag{7.4.10}$$

You might be thinking "Hold up a second! An IRF is defined as having constant velocity, so there can't be an acceleration if we're using special relativity." You'd be kind of right. The observers taking the measurements of this object must be in IRFs (no accelerating), but that doesn't mean the object's rest frame has to be one. It's a highly perpetuated myth that special relativity is incapable of handling accelerations. It just can't handle accelerated reference frames, so as long as we stay out of the object's rest frame, then we're ok.

We can make it look a little more familiar. If we use the chain rule (defined by Eq. 3.1.2) and time dilation (defined by Eq. 7.2.11), then

$$a^{\delta} = \frac{du^{\delta}}{dt}\frac{dt}{d\tau} = \gamma\frac{du^{\delta}}{dt},$$

where γ is defined by Eq. 7.4.4. This γ contains a $\vec{u}$, which is the relative velocity between the object and the frame in which its acceleration is measured. The velocity $\vec{u}$ is *not* the same as the relative velocity $\vec{v}$ between two observers in two different IRFs. The object itself would represent a third frame independent from the other two where it measures its own proper time. It's components in Cartesian coordinates can be shown in matrix notation as

$$a^\delta \longrightarrow \gamma \frac{d}{dt}\begin{bmatrix} u^0 \\ u^1 \\ u^2 \\ u^3 \end{bmatrix} = \gamma \frac{d}{dt}\begin{bmatrix} \gamma c \\ \gamma u_x \\ \gamma u_y \\ \gamma u_z \end{bmatrix} = \begin{bmatrix} \gamma\dot{\gamma} c \\ \gamma\dot{\gamma} u_x + \gamma^2 \dot{u}_x \\ \gamma\dot{\gamma} u_y + \gamma^2 \dot{u}_y \\ \gamma\dot{\gamma} u_z + \gamma^2 \dot{u}_z \end{bmatrix},$$

where the dot accent represents a derivative with respect to coordinate time, d/dt, and we've used the derivative product rule (defined Eq. 3.1.5). We can also write this in shorthand as $\left(\gamma\dot{\gamma}c, \gamma\dot{\gamma}\vec{u} + \gamma^2\dot{\vec{u}}\right)$.

However, we need to get rid of the dots. Let's start with the most difficult dot to remove: $\dot{\gamma}$. It can be evaluated by

$$\dot{\gamma} = \frac{d\gamma}{dt} = \frac{d}{dt}\left[\frac{1}{\sqrt{1 - u^2/c^2}}\right] = \frac{d}{dt}\left[\left(1 - \frac{u^2}{c^2}\right)^{-\frac{1}{2}}\right]$$

$$\dot{\gamma} = -\frac{1}{2}\left(1 - \frac{u^2}{c^2}\right)^{-\frac{3}{2}}\left[-\frac{1}{c^2}\frac{d}{dt}\left(\vec{u} \bullet \vec{u}\right)\right],$$

where we've replaced to u^2 with $\vec{u} \bullet \vec{u}$ for clarity in the next few steps and $\vec{u} = (u_x, u_y, u_z)$ is the coordinate 3-velocity described in Example 7.3.1. Now we're going to use Eq. 4.2.8 on the dot product. If you're not convinced it works for vectors, then use the derivative product rule (defined Eq. 3.1.5) to get

$$\frac{d}{dt}\left(\vec{u} \bullet \vec{u}\right) = \frac{d\vec{u}}{dt} \bullet \vec{u} + \vec{u} \bullet \frac{d\vec{u}}{dt} = 2\vec{u} \bullet \frac{d\vec{u}}{dt}. \tag{7.4.11}$$

This results in

$$\dot{\gamma} = -\frac{1}{2}\left(1 - \frac{u^2}{c^2}\right)^{-\frac{3}{2}}\left[-\frac{2}{c^2}\vec{u} \bullet \frac{d\vec{u}}{dt}\right] = \frac{\gamma^3}{c^2}\left[\vec{u} \bullet \frac{d\vec{u}}{dt}\right],$$

where we've used Eq. 7.4.4 to simplify. In Example 7.3.2, we defined the coordinate 3-acceleration as $\vec{a} = d\vec{u}/dt = \dot{\vec{u}}$, so

$$\dot{\gamma} = \frac{\gamma^3}{c^2}\left(\vec{u} \bullet \vec{a}\right) \tag{7.4.12}$$

and the 4-acceleration becomes

$$a^\delta \longrightarrow \left(\gamma\frac{\gamma^3}{c^2}\left(\vec{u} \bullet \vec{a}\right)c,\ \gamma\frac{\gamma^3}{c^2}\left(\vec{u} \bullet \vec{a}\right)\vec{u} + \gamma^2\vec{a}\right)$$

$$a^\delta \longrightarrow \left(\frac{\gamma^4}{c}\left(\vec{u} \bullet \vec{a}\right),\ \frac{\gamma^4}{c^2}\left(\vec{u} \bullet \vec{a}\right)\vec{u} + \gamma^2\vec{a}\right). \tag{7.4.13}$$

As you can see, this is very complex, but it has very important implications.

The 4-acceleration can be looked at another way. As a proper time derivative of 4-velocity, it represents the rate of change of the world line tangent vector. That makes it the curvature vector to the object's world line (see Figure 7.11). Also, we can take the scalar product of the 4-acceleration with the 4-velocity. Using Eq. 4.2.8 and Eq. 7.4.6 to simplify, The result is

$$u_\delta a^\delta = u_\delta\frac{du^\delta}{d\tau} = \frac{1}{2}\frac{d}{d\tau}\left(u_\delta u^\delta\right) = \frac{1}{2}\frac{d}{d\tau}\left(-c^2\right) = 0, \tag{7.4.14}$$

which is true for all objects in all frames. Since the scalar product is akin to the dot product, this says something about their orthogonality. However, spacetime is a hyperbolic space, so this implies the 4-acceleration and 4-velocity are **hyperbolic orthogonal**. Mathematically, this means something very different than what we normally think of as orthogonal. Physically, since spacetime is hyperbolic and there isn't any other spacetime, hyperbolic orthogonal is the *only* orthogonal. This is one of those "Don't sweat the details" moments.

We can also take the scalar product of the 4-acceleration with itself, but the result isn't as profound as it was for the 4-velocity. We can still use the shorthand notation for the scalar product as we did with the 4-velocity. The general definition of this shorthand is given by

$$T_\delta T^\delta = T_0T^0 + T_1T^1 + T_2T^2 + T_3T^3 = T_0T^0 + \vec{T} \bullet \vec{T},$$

$$T_\delta T^\delta = -\left(T^t\right)^2 + \vec{T} \bullet \vec{T}, \tag{7.4.15}$$

where T^δ is an arbitrary 4-vector and the negative comes from the metric tensor. For the 4-acceleration, this would be

$$a_\delta a^\delta = -\frac{\gamma^8}{c^2}\left(\vec{u}\bullet\vec{a}\right)^2 + \left(\frac{\gamma^4}{c^2}\left(\vec{u}\bullet\vec{a}\right)\vec{u} + \gamma^2\vec{a}\right)\bullet\left(\frac{\gamma^4}{c^2}\left(\vec{u}\bullet\vec{a}\right)\vec{u} + \gamma^2\vec{a}\right)$$

$$a_\delta a^\delta = -\frac{\gamma^8}{c^2}\left(\vec{u}\bullet\vec{a}\right)^2 + \frac{\gamma^8}{c^4}\left(\vec{u}\bullet\vec{a}\right)^2\left(\vec{u}\bullet\vec{u}\right) + \frac{2\gamma^6}{c^2}\left(\vec{u}\bullet\vec{a}\right)^2 + \gamma^4\left(\vec{a}\bullet\vec{a}\right)$$

$$a_\delta a^\delta = \frac{\gamma^6}{c^2}\left(\vec{u}\bullet\vec{a}\right)^2\left[-\gamma^2 + \frac{\gamma^2}{c^2}\left(\vec{u}\bullet\vec{u}\right) + 2\right] + \gamma^4\left(\vec{a}\bullet\vec{a}\right).$$

Since $u^2 = \vec{u}\bullet\vec{u}$ and $a^2 = \vec{a}\bullet\vec{a}$, we get

$$a_\delta a^\delta = \frac{\gamma^6}{c^2}\left(\vec{u}\bullet\vec{a}\right)^2\left[-\gamma^2 + \frac{\gamma^2}{c^2}u^2 + 2\right] + \gamma^4 a^2$$

$$a_\delta a^\delta = \frac{\gamma^6}{c^2}\left(\vec{u}\bullet\vec{a}\right)^2\left[-\gamma^2\left(1 - \frac{u^2}{c^2}\right) + 2\right] + \gamma^4 a^2$$

and, by Eq. 7.4.4,

$$a_\delta a^\delta = \frac{\gamma^6}{c^2}\left(\vec{u}\bullet\vec{a}\right)^2 + \gamma^4 a^2. \tag{7.4.16}$$

This scalar product is still spacetime invariant just like any other real scalar (as opposed to a pseudoscalar), but is not constant like it was for the 4-velocity.

In the rest frame of the object (we'll call it the double-primed frame), we know $\vec{u}'' = 0$ and $\gamma'' = 1$, so the scalar product reduces to

$$a''_\delta a''^\delta = a_p^2. \tag{7.4.17}$$

The quantity a_p is sometimes called the **proper acceleration**, the maximum measurable acceleration. Technically, the rest frame of the object doesn't measure an acceleration since it considers itself to be at rest. That frame actually measures a gravitational force, $F_g = ma_p$, because of the **equivalence principle** (see Section 8.1). The best way to think of proper acceleration is

to imagine it is measured by an IRF that is *momentarily* traveling with the rest frame of the object.

To justify calling it the maximum acceleration, we can solve for coordinate acceleration in terms of proper acceleration by

$$a_p^2 = \frac{\gamma^6}{c^2}\left(\vec{u} \bullet \vec{a}\right)^2 + \gamma^4 a^2,$$

which is just $a_\delta a^\delta = a''_\delta a''^\delta$. Using Eq. 2.2.1 to maintain generality results in

$$a_p^2 = \frac{\gamma^6}{c^2}\left(ua\cos\theta\right)^2 + \gamma^4 a^2$$

$$a_p^2 = \gamma^6 \frac{u^2}{c^2} a^2 \cos^2\theta + \gamma^4 a^2 = \gamma^4\left(\gamma^2\beta^2\cos^2\theta + 1\right)a^2,$$

where $\beta \equiv u/c$. Solving for a, we get

$$a = \frac{a_p}{\gamma^2\sqrt{\gamma^2\beta^2\cos^2\theta + 1}}. \tag{7.4.18}$$

This simplifies in certain special cases given you know θ, the angle between $\vec{u}$ and $\vec{a}$. Since the smallest γ ever gets is one and the smallest β ever gets is zero, the denominator in Eq. 7.4.18 is always greater than or equal to one. Therefore, $a_{\max} = a_p$.

Four-Momentum

If we extend momentum into the 4-vector realm, then it's called **4-momentum** and is defined very similarly to that of 3-momentum. We have

$$\boxed{p^\delta = m_p u^\delta}, \tag{7.4.19}$$

where m_p is the rest mass (or proper mass) and u^δ is the 4-velocity. As long as the rest mass isn't changing, the 4-momentum has all the same properties as the 4-velocity. It's components can be written in shorthand as

$$p^\delta \longrightarrow \left(\gamma m_p c, \gamma m_p \vec{u}\right), \tag{7.4.20}$$

where $\vec{u}$ is the coordinate 3-velocity described in Example 7.3.1 and γ is given by Eq. 7.4.4.

We can easily pick out the **coordinate 3-momentum** as $\vec{p} = m_p\vec{u}$, the 3-momentum that involves a derivative with respect to coordinate time (not proper time). But what's m_pc?! Well, remember the definition of rest energy was $E_p = m_pc^2$? That means $m_pc = E/c$. The time component of the 4-momentum is proportional to the total energy! As a result, it might be more useful to write the 4-momentum's components as

$$p^\delta \longrightarrow \left(\gamma\frac{E_p}{c}, \gamma\vec{p}\right) \tag{7.4.21}$$

or even

$$p^\delta \longrightarrow \left(\frac{E_{\text{rel}}}{c}, \vec{p}_{\text{rel}}\right), \tag{7.4.22}$$

where $E_{\text{rel}} \equiv \gamma E_p$ and $\vec{p}_{\text{rel}} \equiv \gamma\vec{p}$ are defined as the **relativistic energy** and **relativistic 3-momentum**.

This is very convenient because we have incorporated conservation of energy and conservation of 3-momentum into one principle: **conservation of 4-momentum**:

$$p^\delta_{\text{before}} = p^\delta_{\text{after}}, \tag{7.4.23}$$

where either side includes the entire system. The subscripts "before" and "after" refer to measurements taken before and after some event in spacetime. It is also important to distinguish between *conserved* and *invariant* using the following definitions:

- **Spacetime invariant** - A quantity which is the same in all frames.
- **Conserved quantity** - A quantity which is the same before and after an event in a single frame.

E_p is invariant, but not conserved. E_{rel} is conserved, but not invariant. Charge, q, is both conserved and invariant. Do not get these two concepts confused.

We can take the scalar product of the 4-momentum with itself easily by taking advantage Eq. 7.4.6. The result is

$$p_\delta p^\delta = m_p^2 u_\delta u^\delta = -m_p^2c^2, \tag{7.4.24}$$

which is true for all objects, but is only constant if rest mass doesn't change. Upon close inspection, this yields a very familiar and useful invariant equation. Evaluating the scalar product using Eq. 7.4.15, we get

$$-\left(\frac{E_{\text{rel}}}{c}\right)^2 + \vec{p}_{\text{rel}} \bullet \vec{p}_{\text{rel}} = -m_p^2 c^2$$

$$-\left(\frac{E_{\text{rel}}}{c}\right)^2 + (p_{\text{rel}})^2 = -m_p^2 c^2$$

$$-\frac{E_{\text{rel}}^2}{c^2} + p_{\text{rel}}^2 = -m_p^2 c^2$$

$$-E_{\text{rel}}^2 + p_{\text{rel}}^2 c^2 = -m_p^2 c^4$$

$$E_{\text{rel}}^2 = m_p^2 c^4 + p_{\text{rel}}^2 c^2 = E_p^2 + p_{\text{rel}}^2 c^2 \,, \tag{7.4.25}$$

which is often written without the subscripts as $E^2 = m^2 c^4 + p^2 c^2$. However, I find the subscripts help clarify so we don't accidentally substitute in the wrong values.

Four-Force

If we extend net force into the 4-vector realm, then it's called **4-force** and is defined very similarly to that of 3-force. We have

$$F^\delta = \frac{dp^\delta}{d\tau} = m_p a^\delta \,, \tag{7.4.26}$$

where m_p is the rest mass (or proper mass) and a^δ is the 4-acceleration. As long as the rest mass isn't changing, the 4-force has all the same properties as the 4-acceleration. If the rest mass does change, then Eq. 7.4.26 simply has an extra term due to the derivative product rule (defined by Eq. 3.1.5).

The components of the 4-force can be written in shorthand just as we did for the 4-momentum using Eq. 7.4.13. The result is

$$F^\delta \longrightarrow \left(\frac{\gamma^4}{c} m_p \left(\vec{u} \bullet \vec{a}\right), \; \frac{\gamma^4}{c^2} m_p \left(\vec{u} \bullet \vec{a}\right) \vec{u} + \gamma^2 m_p \vec{a}\right), \tag{7.4.27}$$

where $\vec{u}$ is the coordinate 3-velocity described in Example 7.3.1, $\vec{a}$ is the coordinate 3-acceleration described in Example 7.3.2, and γ is given by Eq. 7.4.4. This looks pretty hideous though and it's still assuming constant rest mass. We can write this a little more compactly (and more generally) using

$$F^\delta = \frac{dp^\delta}{d\tau} = \frac{dp^\delta}{dt}\frac{dt}{d\tau} = \gamma\frac{dp^\delta}{dt}$$

and Eq. 7.4.21 to get

$$F^\delta \longrightarrow \gamma\frac{d}{dt}\left(\gamma\frac{E_p}{c}, \gamma\vec{p}\right) = \gamma\frac{d}{dt}\left(\frac{E_{\text{rel}}}{c}, \vec{p}_{\text{rel}}\right)$$

$$\boxed{F^\delta \longrightarrow \left(\gamma\frac{P_{\text{rel}}}{c}, \gamma\vec{F}_{\text{rel}}\right),} \tag{7.4.28}$$

where $\vec{F}_{\text{rel}} \equiv d\vec{p}_{\text{rel}}/dt$ is the **relativistic coordinate 3-force** and $P_{\text{rel}} \equiv dE_{\text{rel}}/dt$ is the **relativistic coordinate power**.

This generalization of net force (essentially Newton's second law) can be used to solve problems in terms of Newton's laws of motion. However, you must use the 4-vector forms of velocity, acceleration, momentum, and force. Newton's first law can be written as

$$\boxed{u^\delta = \frac{dx^\delta}{d\tau} = \text{constant} \quad \left(\text{if } F^\delta = 0\right),} \tag{7.4.29}$$

which looks just like it did in classical physics. Newton's third law of motion does not generalize to special relativity in the sense that we're used to using it. In classical physics, it is consistent to replace the words "action" and "reaction" with the word "force" because they are analogous. This cannot be done if the motion is relativistic because an "action" is a fundamentally unique quantity. Mutual opposite forces are *not necessarily* equal in magnitude. As a result, it is often easier to use Eqs. 7.4.23 and 7.4.25 to solve problems.

Example 7.4.1

A widely used example of special relativity is the decay of a negative pion. A negative pion ($E_{p,\pi} = 139.6$ MeV) is a type of massive particle that often

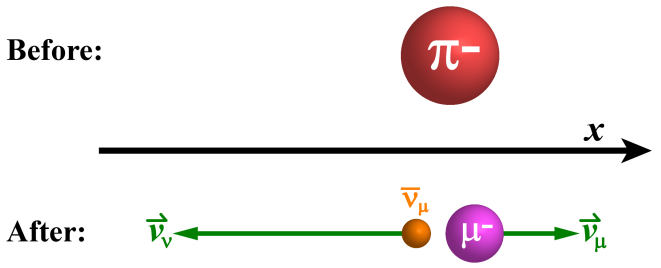


Figure 7.12: This is the before and after picture for the decay of a negative pion into a muon and a muon-antineutrino. It is shown in the rest frame of the pion. The line above the ν indicates the "anti" part of the neutrino.

decays into two other massive particles: a muon ($E_{p,\mu} = 105.7$ MeV) and a muon-antineutrino. Typically, a neutrino (designated by the symbol ν) is considered massless because it is very small compared to other particles (e.g. $E_{p,\nu} \ll E_{p,\mu}$), but it is not actually massless. We're not quite prepared to deal with massless particles, but unfortunately the neutrino's mass is only approximately known. For the purposes of this example, we'll go with a "middle of the road" estimate of $E_{p,\nu} = 1.5$ eV (*not* MeV).

Let's start this problem by stating that all measurements will be taken in the pion's rest frame. Now we'll apply conservation of 4-momentum (given by Eq. 7.4.23) using Figure 7.12, which results in

$$p_\pi^\delta = p_\mu^\delta + p_\nu^\delta,$$

where δ is a free index capable of taking on four different values (Note that π, μ, and ν are not indices but just labels for the particles). This is actually four equations: one for each component of 4-momentum. We can write these component equations out using Eq. 7.4.21 in shorthand notation as

$$\left(\frac{E_{p,\pi}}{c}, 0\right) = \left(\gamma_\mu \frac{E_{p,\mu}}{c}, \gamma_\mu p_\mu \hat{x}\right) + \left(\gamma_\nu \frac{E_{p,\nu}}{c}, -\gamma_\nu p_\nu \hat{x}\right)$$

or in matrix notation as

$$\begin{bmatrix} E_{p,\pi}/c \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \gamma_\mu E_{p,\mu}/c \\ \gamma_\mu p_\mu \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \gamma_\nu E_{p,\nu}/c \\ -\gamma_\nu p_\nu \\ 0 \\ 0 \end{bmatrix}.$$

where, in both cases, we have taken $\gamma_\pi = 1$ because $u_\pi = 0$ in it's own rest frame.

Either way you do it, you still end up with two equations. Performing the addition and multiplying all components by c, we get

$$\begin{aligned} E_{p,\pi} &= \gamma_\mu E_{p,\mu} + \gamma_\nu E_{p,\nu} && \text{(7.4.30a)} \\ 0 &= \gamma_\mu p_\mu c - \gamma_\nu p_\nu c && \text{(7.4.30b)} \end{aligned}$$

where the y and z components are unnecessary. However, Eq. Set 7.4.30 has two equations, but four unknowns. We need two other equations to solve this system and they will come from Eq. 7.4.25. With a little manipulation, we get

$$\gamma^2 E_p^2 = E_p^2 + \gamma^2 p^2 c^2 \quad \Rightarrow \gamma p c = \sqrt{\gamma^2 - 1}\, E_p,$$

which we can use on both the muon and the neutrino. This gives

$$\begin{aligned} \gamma_\mu p_\mu c &= \sqrt{\gamma_\mu^2 - 1}\, E_{p,\mu} && \text{(7.4.31a)} \\ \gamma_\nu p_\nu c &= \sqrt{\gamma_\nu^2 - 1}\, E_{p,\nu} && \text{(7.4.31b)} \end{aligned}$$

where we've used the appropriate subscripts.

We can focus for now on finding the two gamma factors: γ_μ and γ_ν. Eq. 7.4.30b yeilds

$$\gamma_\mu p_\mu c = \gamma_\nu p_\nu c$$

and, with Eq. Set 7.4.31,

$$\sqrt{\gamma_\mu^2 - 1}\, E_{p,\mu} = \sqrt{\gamma_\nu^2 - 1}\, E_{p,\nu}$$

$$\left(\gamma_\mu^2 - 1\right) E_{p,\mu}^2 = \left(\gamma_\nu^2 - 1\right) E_{p,\nu}^2$$

$$\gamma_\mu^2 E_{p,\mu}^2 - E_{p,\mu}^2 = \gamma_\nu^2 E_{p,\nu}^2 - E_{p,\nu}^2.$$

Using Eq. 7.4.30a to solve for $\gamma_\nu E_{p,\nu}$ yields $\gamma_\nu E_{p,\nu} = E_{p,\pi} - \gamma_\mu E_{p,\mu}$, so

$$\gamma_\mu^2 E_{p,\mu}^2 - E_{p,\mu}^2 = \left(E_{p,\pi} - \gamma_\mu E_{p,\mu}\right)^2 - E_{p,\nu}^2$$

$$\gamma_\mu^2 E_{p,\mu}^2 - E_{p,\mu}^2 = E_{p,\pi}^2 - 2\gamma_\mu E_{p,\mu} E_{p,\pi} + \gamma_\mu^2 E_{p,\mu}^2 - E_{p,\nu}^2.$$

If we cancel and group like terms, then we get

$$-E_{p,\mu}^2 = E_{p,\pi}^2 - 2\gamma_\mu E_{p,\mu} E_{p,\pi} - E_{p,\nu}^2$$

$$2\gamma_\mu E_{p,\mu} E_{p,\pi} = E_{p,\pi}^2 + E_{p,\mu}^2 - E_{p,\nu}^2$$

$$\gamma_\mu = \frac{E_{p,\pi}^2 + E_{p,\mu}^2 - E_{p,\nu}^2}{2E_{p,\mu} E_{p,\pi}}. \tag{7.4.32}$$

Substituting in all the rest energies gives a value of $\gamma_\mu = 1.039$ for our example. This being close to a value of one implies that the muon is moving relatively slow after the decay.

We can now summarize by finding all there is to know about the muon: Eq. 7.4.4 gives us the speed, $\gamma_\mu E_{p,\mu}$ is the relativistic energy, $\gamma_\mu E_{p,\mu} - E_{p,\mu}$ is the kinetic energy, and Eq. 7.4.31a gives us the relativistic momentum. Therefore,

$$\left\{ \begin{array}{rcl} \gamma_\mu & = & 1.039 \\ u_\mu & = & 0.271c \\ p_\mu^\delta & \longrightarrow & \left(109.8 \, \dfrac{\text{MeV}}{c}, \; 29.78 \, \dfrac{\text{MeV}}{c} \hat{x} \right) \\ KE_\mu & = & 4.1 \, \text{MeV} \end{array} \right\},$$

but what about the neutrino?

It is evident from the numerator in Eq. 7.4.32 that $E_{p,\nu}^2$ makes a negligible contribution to the values in this example. That in mind, we expect γ_ν to be very large and u_ν (pardon the pun) to be very nearly c. If we use Eq. 7.4.30a, we get

$$\gamma_\nu E_{p,\nu} = E_{p,\pi} - \gamma_\mu E_{p,\mu}$$

$$\gamma_\nu = \frac{E_{p,\pi} - \gamma_\mu E_{p,\mu}}{E_{p,\nu}},$$

which gives a value of $\gamma_\nu = 1.99 \times 10^7$ corresponding to a speed of $u_\nu = 0.999\,999\,999\,999\,999\,c$. That's 15 nines after the decimal point! We can now

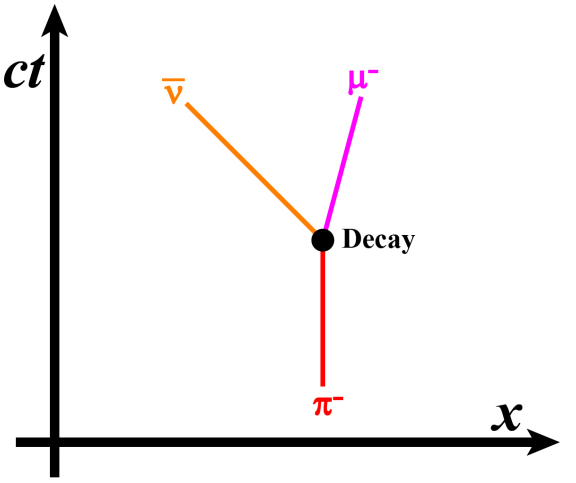


Figure 7.13: This is a spacetime diagram showing the decay of a negative pion into a muon and a muon-antineutrino. The coordinate system shown is the rest frame of the pion. It is clear that the antineutrino zips off *almost* along a light-like world line due to its very low mass.

summarize by finding all there is to know about the neutrino: Eq. 7.4.4 gives us the speed, $\gamma_\nu E_{p,\nu}$ is the relativistic energy, $\gamma_\nu E_{p,\nu} - E_{p,\nu}$ is the kinetic energy, and Eq. 7.4.31b gives us the relativistic momentum. Therefore,

$$\left\{ \begin{array}{rcl} \gamma_\nu & = & 1.99 \times 10^7 \\ u_\nu & = & 0.999\,999\,999\,999\,999\,c \\ p_\nu^\delta & \longrightarrow & \left(29.8\,\dfrac{\text{MeV}}{c},\ -29.8\,\dfrac{\text{MeV}}{c}\hat{x} \right) \\ KE_\nu & = & 29.8\,\text{MeV} \end{array} \right\},$$

where it can be seen that the neutrino's total energy is entirely kinetic energy within the significant figures we've kept. It's also apparent from the neutrino's 4-momentum that it's traveling on *very nearly* a null world line since its time and space components are the same (See Figure 7.13).

Just as a check, you can add the 4-momentums of the muon and the muon-antineutrino and you'll arrive at the 4-momentum of the pion. We can take note again that rest energy, E_p, is *not* conserved (as expected) since $139.6\,\text{MeV} \neq 105.7\,\text{MeV} + 1.5\,\text{eV}$, but *is* invariant since each of those three measurements is the same in all frames of reference. The missing 33.9

MeV went into the kinetic energy (i.e. the motion) of the muon and muon-antineutrino. Rest energy isn't really anything new. For example, the pion is made of more fundamental particles, so the 139.6 MeV is simply the kinetic energy of those particles (the pion's rest frame is the center of mass frame for those particles) plus the potential energy between those particles (i.e. the nuclear bonds).

The relativistic energy, E_{rel}, *is* conserved since $139.6\,\text{MeV} = 109.8\,\text{MeV} + 29.8\,\text{MeV}$ in the rest frame of the pion and $145.0\text{MeV} = 105.7\text{MeV} + 39.3\text{MeV}$ in the rest frame of the muon. However, relativistic energy is *not* invariant since $E_{\text{rel},\mu} = 109.8$ MeV in the rest frame of the pion ($\gamma_\mu = 1.039$), but $E_{\text{rel},\mu} = 105.7$ MeV in the rest frame of the muon ($\gamma_\mu = 1$). The same can be shown for the pion ($E_{\text{rel},\pi} = 145.0$ MeV) and the muon-antineutrino ($E_{\text{rel},\nu} = 39.3$ MeV). The pion has the extra 5.4 MeV due to its motion in the muon's rest frame.

Total charge, on the other hand, is $q = -1.602 \times 10^{-19}$ C before and after the decay, which makes it conserved. It is also measured to be $q = -1.602 \times 10^{-19}$ C in every frame of reference, which makes it invariant. This is a very unique quality of charge and is very important in all particle decays.

7.5 Relativistic Electrodynamics

If we want to formulate electrodynamics under the premise of spacetime, then we'll need to write all the quantities in electrodynamics as 4-vectors (or at least 4-pseudovectors). The covariant derivative described in Section 6.7 will help us with this process. In Cartesian coordinates (which is what we tend to stick with in special relativity) for an arbitrary 4-vector T^δ, it is

$$\nabla_\alpha T^\delta = \frac{\partial}{\partial x^\alpha} T^\delta = \frac{\partial T^\delta}{\partial x^\alpha},$$

where upper indices in the denominator of a derivative are actually lower indices. If we want a scalar result, then

$$\nabla_\alpha T^\alpha = \frac{\partial T^\alpha}{\partial x^\alpha} = \frac{\partial T^0}{\partial x^0} + \frac{\partial T^1}{\partial x^1} + \frac{\partial T^2}{\partial x^2} + \frac{\partial T^3}{\partial x^3},$$

where α has become a summation index. Using a shorthand similar to Eq. 7.4.15, we can write this as

$$\nabla_\alpha T^\alpha = \frac{1}{c}\frac{\partial T^t}{\partial t} + \vec{\nabla} \bullet \vec{T}, \tag{7.5.1}$$

where $\vec{\nabla}$ is the three-dimensional del operator, $\vec{T}$ is the spatial 3-vector, and we've used $x^0 = ct$.

Looking at charge continuity (defined by Eq. 5.3.22), we see that it involves both the charge density ρ and the current density $\vec{J}$. It also fits the form of Eq. 7.5.1. A little rearranging gives us

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \bullet \vec{J} = 0$$

$$\frac{1}{c}\frac{\partial (c\rho)}{\partial t} + \vec{\nabla} \bullet \vec{J} = 0,$$

where we can say $c\rho$ represents the time component of the current density 4-vector (or **4-current**). Therefore, in shorthand notation, we can write the 4-current as

$$J^\alpha \longrightarrow \left(c\rho, \vec{J}\right), \tag{7.5.2}$$

where ρ and $\vec{J}$ are considered relativistic quantities. In terms of the 4-velocity of the charges, this is

$$J^\alpha = \rho_p u^\alpha \longrightarrow (\gamma \rho_p c, \gamma \rho_p \vec{u}), \tag{7.5.3}$$

where ρ_p is the **proper charge density** (or minimum measurable charge density) measured in the rest frame of the charge. Charge may be invariant in spacetime, but charge density involves volume, one dimension of which experiences length contraction. Using Eq. 7.5.2, we can write the charge continuity equation as

$$\nabla_\alpha J^\alpha = 0, \tag{7.5.4}$$

where ∇ is the covariant derivative.

The same can be done for electric potential ϕ and magnetic potential $\vec{A}$, but we have to be a little more careful. We'll be using the Lorenz gauge (not

to be confused with Lorentz), given by Eq. 5.6.13. A little rearranging gives us

$$\frac{1}{c^2}\frac{\partial \phi}{\partial t} + \vec{\nabla} \bullet \vec{A} = 0$$

$$\frac{1}{c}\frac{\partial \left(\phi/c\right)}{\partial t} + \vec{\nabla} \bullet \vec{A} = 0,$$

where we can say ϕ/c represents the time component of the potential 4-vector (or **4-potential**). Therefore, in shorthand notation, we can write the 4-current as

$$A^\alpha \longrightarrow \left(\frac{\phi}{c}, \vec{A}\right). \tag{7.5.5}$$

This means we can now write the Lorenz gauge as

$$\nabla_\alpha A^\alpha = 0\,, \tag{7.5.6}$$

where ∇ is the covariant derivative. However, Eqs. 7.5.5 and 7.5.6 don't work under any gauges other than the Lorenz gauge. Conveniently, this is the gauge we used to derive Maxwell's equations in Section 5.6.

Maxwell's Equations with Potentials

We'll keep things short by starting with Eq. 5.6.16. A little rearranging gives us

$$-\frac{1}{c^2}\frac{\partial^2 \vec{A}}{\partial t^2} + \vec{\nabla}^2 \vec{A} = -\mu_0 \vec{J},$$

which involves second derivatives. We can define a second derivative operator called the **d'Alembertian** given by

$$\Box \equiv \nabla_\delta \nabla^\delta = g^{\mu\delta} \nabla_\delta \nabla_\mu,$$

which, using Eq. 7.4.15, becomes

$$\Box \equiv -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \vec{\nabla}^2\,. \tag{7.5.7}$$

Now we get

$$\Box\vec{A} = -\mu_0\vec{J},$$

which looks much simpler. Note here that $\vec{A}$ and $\vec{J}$ are the spatial components of their respective 4-vector counterparts.

Using Eq. 5.6.15, we can get a very similar result. A little rearranging gives us

$$-\frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} + \vec{\nabla}^2\phi = -\frac{\rho}{\epsilon_0}.$$

If we multiple through by $c/c^2 = c\mu_0\epsilon_0$ (we used Eq. 5.5.4), then we get

$$-\frac{1}{c^2}\frac{\partial^2(\phi/c)}{\partial t^2} + \vec{\nabla}^2\left(\frac{\phi}{c}\right) = -\mu_0(c\rho)$$

$$-\frac{1}{c^2}\frac{\partial^2(\phi/c)}{\partial t^2} + \vec{\nabla}^2\left(\frac{\phi}{c}\right) = -\mu_0(c\rho)$$

$$\Box\left(\frac{\phi}{c}\right) = -\mu_0(c\rho).$$

The parenthetical quantities on both sides just represent time components of the 4-potential and 4-current, respectfully. Therefore, we can conclude in general that

$$\Box A^\alpha = -\mu_0 J^\alpha, \tag{7.5.8}$$

where we've simplified Maxwell's equations down to one equation using tensor analysis.

Electromagnetic Field Tensor

Eq. 7.5.8 is extremely elegant in that it is only one equation. However, you might want to use fields rather than potentials given a particular situation. Writing the electric field $\vec{E}$ and the magnetic field $\vec{B}$ in spacetime is much trickier than it was for the potential functions. As was suggested in Section 5.7, there is no real distinction between $\vec{E}$ and $\vec{B}$. They tend to blur together

into what we called the electromagnetic field: $\vec{E} + \vec{v} \times \vec{B}$. You can have one, or the other, or both depending on which IRF you're observing from, so it stands to reason they are really just one quantity.

When it came to 4-current and 4-potential, we were merging a scalar with a vector resulting in four components. To make an **electromagnetic field tensor**, we need to merge two vectors together. That's a total of six components, not four (two too many to be a 4-vector). The next 4-quantity available with more components is a rank-2 tensor. This has 16 components and we only need six, which is something we'll need to address. We'd also like whatever this is to be a real tensor rather than a pseudotensor, so it will obey the Lorentz transformation.

We already know by Eqs. 5.6.1 and 5.6.2 that the fields can be defined in terms of the potentials. We also know the 4-potential is given by Eq. 7.5.5. Combining Eqs. 5.6.1 and 7.5.5, we get

$$\vec{E} = -\vec{\nabla}\phi - \frac{\partial \vec{A}}{\partial t} = -\vec{\nabla}\left(cA^t\right) - \frac{\partial \vec{A}}{\partial t}$$

$$-\frac{\vec{E}}{c} = \vec{\nabla}A^t + \frac{1}{c}\frac{\partial \vec{A}}{\partial t},$$

where we've multiplied through by $-1/c$. It sort of looks like a covariant derivative on the right side, but not quite since the components are mixed. Let's get a better look at this through its components, which are given by

$$\left\{\begin{array}{rcl} -\dfrac{E^x}{c} & = & \nabla^x A^t + \dfrac{1}{c}\dfrac{\partial A^x}{\partial t} \\ -\dfrac{E^y}{c} & = & \nabla^y A^t + \dfrac{1}{c}\dfrac{\partial A^y}{\partial t} \\ -\dfrac{E^z}{c} & = & \nabla^z A^t + \dfrac{1}{c}\dfrac{\partial A^z}{\partial t} \end{array}\right\}$$

$$\left\{\begin{array}{rcl} -E^x/c & = & \nabla^x A^t - \nabla^t A^x \\ -E^y/c & = & \nabla^y A^t - \nabla^t A^y \\ -E^z/c & = & \nabla^z A^t - \nabla^t A^z \end{array}\right\}, \tag{7.5.9}$$

where we've used

$$\nabla^\delta = g^{\mu\delta}\nabla_\mu$$

$$\nabla^t = g^{tt}\nabla_t + g^{xt}\nabla_x + g^{yt}\nabla_y + g^{zt}\nabla_z = -\nabla_t$$

to keep the same derivative throughout (in this case, the contravariant derivative).

We can also perform this same process for Eq. 5.6.2. In index notation for dimension-3, the magnetic field can be written

$$B_i = \varepsilon_{ijk}\nabla^j A^k,$$

where ε_{ijk} is the dimension-3 Levi-Civita tensor defined by Eq. 6.6.4. Since all three indices must be different, this leaves us with the components of

$$\left\{\begin{array}{rcl} B^x &=& \nabla^y A^z - \nabla^z A^y \\ B^y &=& \nabla^z A^x - \nabla^x A^z \\ B^z &=& \nabla^x A^y - \nabla^y A^x \end{array}\right\}, \tag{7.5.10}$$

where we've realized $B_i = B^i$ in Cartesian 3-space due to Eq. 6.4.5. These have exactly the same form as the electric field components did in Eq. 7.5.9.

Let's take advantage of this pattern and define the contravariant electromagnetic field tensor to be

$$\mathcal{F}^{\alpha\delta} = \nabla^\alpha A^\delta - \nabla^\delta A^\alpha. \tag{7.5.11}$$

This represents an antisymmetric dimension-4 rank-2 tensor. As a dimension-4 rank-2 tensor, it has the expected $4^2 = 16$ components. Since it's antisymmetric (i.e. $\mathcal{F}^{\alpha\delta} = -\mathcal{F}^{\delta\alpha}$), the diagonal components must be zero leaving only 12 components, but half of those are just opposite-sign duplicates. That's six independent components!

Using Eqs. 7.5.9 and 7.5.10 with Eq. 7.5.11, we get a contravariant form of

$$\mathcal{F}^{\alpha\delta} \longrightarrow \begin{bmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & B_z & -B_y \\ -E_y/c & -B_z & 0 & B_x \\ -E_z/c & B_y & -B_x & 0 \end{bmatrix} \tag{7.5.12}$$

and a covariant form of

$$\mathcal{F}_{\alpha\delta} \longrightarrow \begin{bmatrix} 0 & -E_x/c & -E_y/c & -E_z/c \\ E_x/c & 0 & B_z & -B_y \\ E_y/c & -B_z & 0 & B_x \\ E_z/c & B_y & -B_x & 0 \end{bmatrix}. \tag{7.5.13}$$

It's only the electric field components that change from contravariant to covariant because they're the only components to have an index value of zero (or a value of t depending on how you look at it), which is the index that experiences sign change according to Eq. 7.3.8. This is a real tensor as it transforms according to

$$\mathcal{F}'^{\mu\nu} = \Lambda^{\mu}_{\alpha}\Lambda^{\nu}_{\delta}\mathcal{F}^{\alpha\delta}, \tag{7.5.14}$$

where we need a Lorentz transformation matrix for each index.

The scalar product of the EMF tensor with itself is a spacetime invariant as we'd expect. It takes the form

$$\mathcal{F}_{\alpha\delta}\mathcal{F}^{\alpha\delta} = 2\frac{E_x E_x}{c^2} + 2\frac{E_y E_y}{c^2} + 2\frac{E_z E_z}{c^2} - 2B_x B_x - 2B_y B_y - 2B_z B_z$$

$$\mathcal{F}_{\alpha\delta}\mathcal{F}^{\alpha\delta} = 2\left(\frac{\vec{E}\bullet\vec{E}}{c^2} - \vec{B}\bullet\vec{B}\right), \tag{7.5.15}$$

where α and δ are both summation indices (equivalent to taking the trace of the matrix product). The determinant of the tensor,

$$\det(\mathcal{F}) = \frac{E_x B_x + E_y B_y + E_z B_z}{c^2} = \frac{\vec{E}\bullet\vec{B}}{c^2}, \tag{7.5.16}$$

is also spacetime invariant. Even if the electric and magnetic field components change between frames, the results of Eqs. 7.5.15 and 7.5.16 will not.

Example 7.5.1

In your IRF, a charge q (which is invariant) is moving to the right with a constant speed of u (which is *not* invariant). Determine the E-field at an arbitrary point around this moving charge.

- Let's start in the charge's rest frame (double-primed frame in Figure 7.14) where we know exactly what the electric field looks like. It's given exactly by Coulomb's law (defined by Eq. 5.2.5),

$$\vec{E} = k_E\frac{q}{(r'')^2}\hat{r}'' = k_E\frac{q}{\left|\vec{r}''_p - \vec{r}''_q\right|^3}\left(\vec{r}''_p - \vec{r}''_q\right)$$

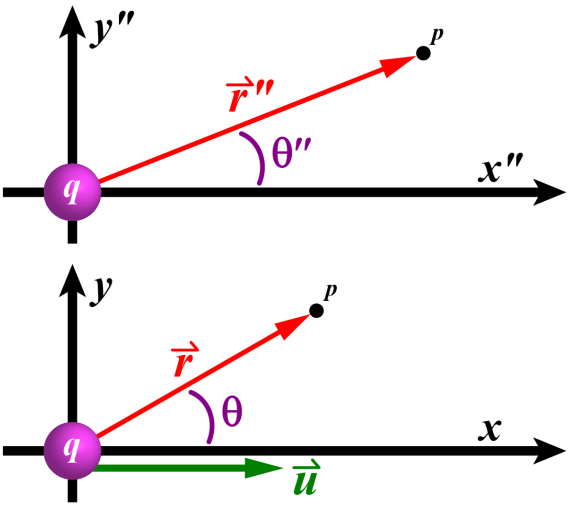


Figure 7.14: There are two IRFs shown: the charge's rest frame (double-primed) and another frame (unprimed) in which it is moving with constant velocity $\vec{u}$ in the x-direction. A position vector of an arbitrary point p is also shown for both frames. This point p appears closer to the charge along the direction of motion in the unprimed frame due to length contraction.

where $\vec{r}''_p$ is the location of the arbitrary point and $\vec{r}''_q$ is the location of the charge in the rest frame of the charge. For simplicity, since it's the only object in the system, we can put the charge at the origin (i.e. $\vec{r}''_q = 0$ and $\vec{r}''_p = \vec{r}''$) allowing us to drop the more complicated notation. The result in Cartesian coordinates is

$$\vec{E} = k_E \frac{q}{(r'')^3}\vec{r}'' = \frac{k_E q}{\left[(x'')^2 + (y'')^2 + (z'')^2\right]^{\frac{3}{2}}}\left(x''\hat{x} + y''\hat{y} + z''\hat{z}\right).$$

- Before we can transform from this rest frame out to an arbitrary IRF (just as we did in Eq. 7.4.7), we need to simply a bit. We'll be using the standard Lorentz transformation matrix from this chapter which assumes the relative motion between the two frames is only in the x-direction. Since we're starting from the rest frame, this implies the motion of the charge in new frame should be measured in the x-direction. Just as in Example 5.2.1, this will result in cylindrical symmetry about the x-axis and, therefore, (x'', s'', ϕ'') as a set of generalized coordinates.

The electric field in the rest frame of the charge can now be written as

$$\vec{E} = \frac{k_E q}{\left[(x'')^2 + (y'')^2\right]^{\frac{3}{2}}} \left(x''\hat{x} + y''\hat{y}\right),$$

where we have suppressed the z-direction making $y'' = s''$ (we're staying in the xy-plane and we'll bring back the z-direction later).

- Now we need to write out this electric field in the form of the EMF tensor since it wont obey Lorentz transformations otherwise. As usual, it's more convenient to work with contravariant forms, so we'll use Eq. 7.5.12 to get

$$\mathcal{F}''^{\alpha\delta} \longrightarrow \frac{k_E q}{c\left[(x'')^2 + (y'')^2\right]^{\frac{3}{2}}} \begin{bmatrix} 0 & x'' & y'' & 0 \\ -x'' & 0 & 0 & 0 \\ -y'' & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where all the magnetic field components are zero because stationary charges don't generate magnetic fields. We've also pulled out all quantities common to all non-zero components.

- The transformation equation is given in index notation by Eq. 7.5.14, but some of us may still feel more comfortable with matrix notation. If we intend to write this transformation in terms of matrix multiplication, then we need to be more careful. Because matrices do not commute, the order will matter. We need to make sure we're summing over columns in the first matrix and rows in the second. A little rearranging gives

$$\mathcal{F}^{\mu\nu} = \Lambda^{\mu}_{\alpha} \mathcal{F}''^{\alpha\delta} \Lambda^{\nu}_{\delta},$$

where we define the first index on $\mathcal{F}$ as the *row* index and the second as the *column* index (Λ is symmetric so it doesn't matter which index is which). Let's do this step-by-step so we don't get lost. Starting with the last two matrices, we get

$$\mathcal{F}''^{\alpha\delta}\Lambda^{\nu}_{\delta} \longrightarrow \frac{k_E q}{c\left[(x'')^2 + (y'')^2\right]^{\frac{3}{2}}} \begin{bmatrix} 0 & x'' & y'' & 0 \\ -x'' & 0 & 0 & 0 \\ -y'' & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathcal{F}''^{\alpha\delta}\Lambda^{\nu}_{\delta} \longrightarrow \frac{k_E q}{c\left[(x'')^2+(y'')^2\right]^{\frac{3}{2}}}\begin{bmatrix}\gamma\beta x'' & \gamma x'' & y'' & 0\\ -\gamma x'' & -\gamma\beta x'' & 0 & 0\\ -\gamma y'' & -\gamma\beta y'' & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

To get the final result, we just multiply another Λ on the front which gives

$$\mathcal{F}^{\mu\nu} \longrightarrow \frac{k_E q}{c\left[(x'')^2+(y'')^2\right]^{\frac{3}{2}}}\begin{bmatrix}\gamma & \gamma\beta & 0 & 0\\ \gamma\beta & \gamma & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\gamma\beta x'' & \gamma x'' & y'' & 0\\ -\gamma x'' & -\gamma\beta x'' & 0 & 0\\ -\gamma y'' & -\gamma\beta y'' & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

$$\mathcal{F}^{\mu\nu} \longrightarrow \frac{k_E q}{c\left[(x'')^2+(y'')^2\right]^{\frac{3}{2}}}\begin{bmatrix}0 & \gamma^2(1-\beta^2)x'' & \gamma y'' & 0\\ -\gamma^2(1-\beta^2)x'' & 0 & \gamma\beta y'' & 0\\ -\gamma y'' & -\gamma\beta y'' & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

and, by the definition of γ (Eq. 7.2.9),

$$\mathcal{F}^{\mu\nu} \longrightarrow \frac{k_E q}{c\left[(x'')^2+(y'')^2\right]^{\frac{3}{2}}}\begin{bmatrix}0 & x'' & \gamma y'' & 0\\ -x'' & 0 & \gamma\beta y'' & 0\\ -\gamma y'' & -\gamma\beta y'' & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}.$$

- Unfortunately, we still have some double-primes lingering around from the rest frame. We already know the components of spacetime position can be different depending on which IRF is taking the measurements. A length contraction is witnessed along the direction of motion and, in this case,

$$\frac{x''}{\gamma} = x \quad \Rightarrow x'' = \gamma x$$

from Eq. 7.2.12 and $y'' = y$. Therefore,

$$\mathcal{F}^{\mu\nu} \longrightarrow \frac{k_E q}{c\left(\gamma^2x^2+y^2\right)^{\frac{3}{2}}}\begin{bmatrix}0 & \gamma x & \gamma y & 0\\ -\gamma x & 0 & \gamma\beta y & 0\\ -\gamma y & -\gamma\beta y & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

$$\mathcal{F}^{\mu\nu} \longrightarrow \frac{\gamma k_E q}{c\left[\gamma^2 x^2 + y^2\right]^{\frac{3}{2}}} \begin{bmatrix} 0 & x & y & 0 \\ -x & 0 & \beta y & 0 \\ -y & -\beta y & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{7.5.17}$$

- This actually gives us two results. First, by Eq. 7.5.12, we get an electric field in the new (arbitrary frame) of

$$\vec{E} = \frac{\gamma k_E q}{\left(\gamma^2 x^2 + y^2\right)^{\frac{3}{2}}} \left(x\hat{x} + y\hat{y}\right),$$

or, better yet,

$$\boxed{\vec{E} = \frac{\gamma k_E q}{\left(\gamma^2 x^2 + s^2\right)^{\frac{3}{2}}} \left(x\hat{x} + s\hat{s}\right)}, \tag{7.5.18}$$

where $s = \sqrt{y^2 + z^2}$ is defined in our version of cylindrical coordinates given by (x, s, ϕ). We should take note that the factor of c in the denominator has disappeared because the EMF tensor components include it for $\vec{E}$.

We might get a better feel for what this field looks like if we use the angle θ in Figure 7.14 to generalize. We know $\vec{r} = x\hat{x} + s\hat{s}$ as well as

$$\left\{ \begin{matrix} x = r\cos\theta \\ s = r\sin\theta \end{matrix} \right\}$$

where θ is the angle between $\vec{u}$ and $\vec{r}$. Now the E-field can be written

$$\vec{E} = \frac{\gamma k_E q}{\left(\gamma^2 r^2 \cos^2\theta + r^2 \sin^2\theta\right)^{\frac{3}{2}}} \vec{r}$$

$$\boxed{\vec{E} = \frac{\gamma}{\left(\gamma^2 \cos^2\theta + \sin^2\theta\right)^{\frac{3}{2}}} k_E \frac{q}{r^3} \vec{r}}, \tag{7.5.19}$$

which looks a lot like Coulomb's law (defined by Eq. 5.2.5) but with a hideous factor out front. This is the generalized form of the electric field at an arbitrary point around a moving point charge. Along the

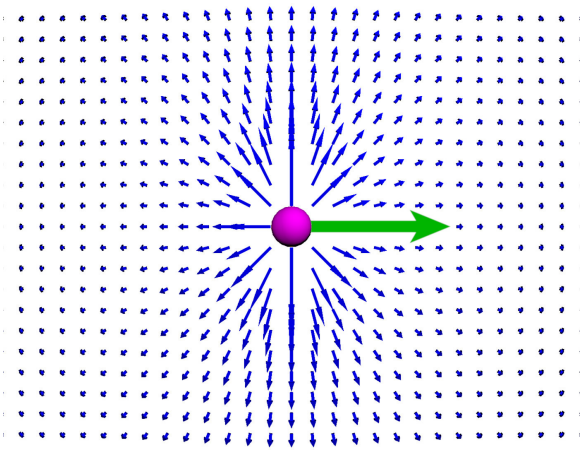


Figure 7.15: This the E-field surrounding a charge moving in the positive x-direction (horizontal in the figure) at a constant speed of $u = 0.7c$. You can see very clearly the compression along the horizontal and the expansion along the vertical.

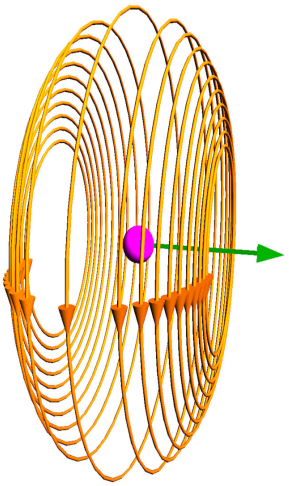


Figure 7.16: This is the B-field surrounding a charge moving in the positive x-direction (velocity shown by a green arrow) at a constant speed of $u = 0.7c$. Each line is of equal $\vec{B}$ and it is evident how quickly the B-field drops off as the points in question are further away from the charge.

direction of motion (i.e. $\theta = 0$), the hideous factor reduces to $1/\gamma^2$ implying the electric field is less than we'd expect from Coulomb's law. Orthogonal to the direction of motion (i.e. $\theta = 90°$), the hideous factor reduces to γ implying the electric field is greater than we'd expect from Coulomb's law. This is shown visually in Figure 7.15.

- The other result from Eq. 7.5.17 is that, in the new frame, there is also a magnetic field. This makes sense since the charge is moving in the new frame, but the beautiful thing about this is that we didn't even have to think about it. It appeared automatically! This is an example of the completeness that comes with the EMF tensor. By Eq. 7.5.12, this relativistic magnetic field is

$$\vec{B} = \frac{\gamma k_E q}{c\left(\gamma^2 x^2 + y^2\right)^{\frac{3}{2}}} \beta y \hat{z}$$

$$\vec{B} = \frac{\gamma k_M q u y}{\left(\gamma^2 x^2 + y^2\right)^{\frac{3}{2}}} \hat{z},$$

where we used $\beta \equiv u/c$ and $k_M = k_E/c^2$. Just as we did with the E-field, if we write this under the generalized coordinates (x, s, ϕ), then

$$\boxed{\vec{B} = \frac{\gamma k_M q u s}{\left(\gamma^2 x^2 + s^2\right)^{\frac{3}{2}}} \hat{\phi}}, \tag{7.5.20}$$

where $\hat{z}$ is just $\hat{\phi}$ in the xy-plane (where all our original math took place).

We might get a better feel for what this field looks like if we use the angle θ in Figure 7.14 to generalize. We know $\vec{r} = x\hat{x} + s\hat{s}$ as well as

$$\left\{ \begin{array}{l} x = r\cos\theta \\ s = r\sin\theta \end{array} \right\}$$

where θ is the angle between $\vec{u}$ and $\vec{r}$. Now the B-field can be written

$$\vec{B} = \frac{\gamma k_M q u r \sin\theta}{\left(\gamma^2 r^2 \cos^2\theta + r^2 \sin^2\theta\right)^{\frac{3}{2}}} \hat{\phi}$$

$$\boxed{\vec{B} = \frac{\gamma \sin\theta}{\left(\gamma^2 \cos^2\theta + \sin^2\theta\right)^{\frac{3}{2}}} k_M \frac{qu}{r^2}\hat{\phi}}, \tag{7.5.21}$$

where $\hat{\phi}$ is defined counterclockwise about the x-axis as viewed from positive infinity. This is the generalized form of the magnetic field at an arbitrary point around a moving point charge. Along the direction of motion (i.e. $\theta = 0$), the hideous factor reduces to zero implying the magnetic field is zero along this axis (the x-axis in our example). Orthogonal to the direction of motion (i.e. $\theta = 90°$), the hideous factor reduces to γ implying the magnetic field is stronger further from the charge in that direction. This is shown visually in Figure 7.16.

Example 7.5.2

In one IRF, we observe that two equal positive charges ($q_1 = q_2 = q$ which is invariant) are moving in opposite directions with equal constant speed ($u_1 = u_2 = u$ which is *not* invariant) as shown in Figure 7.17. At closest approach, these charges are separated by a distance R, which does *not* experience length contraction since it's orthogonal to the motion of both charges. Determine the Lorentz force on q_2 due to q_1 (i.e. $\vec{F}_{21}$) in this frame at closest approach. Also, determine the same Lorentz force in the rest frame of q_1 and the rest frame of q_2.

- In the IRF described by this example, the E-field generated by q_1 at the location of q_2 is given by Eq. 7.5.19 to be

$$\vec{E}_1 = \gamma_1 \, k_E \frac{q_1}{r^3}\vec{r} = \gamma_1 \, k_E \frac{q}{R^2}\hat{y}$$

because $\theta = 90°$, $q_1 = q$, $r = R$, and $\vec{r} = R\hat{y}$. By the same logic, the B-field generated by q_1 at the location of q_2 is given by Eq. 7.5.21 to be

$$\vec{B}_1 = \gamma_1 \, k_M \frac{qu_1}{r^2}\hat{\phi} = \gamma_1 \, \frac{k_E}{c^2}\frac{qu}{R^2}\hat{z},$$

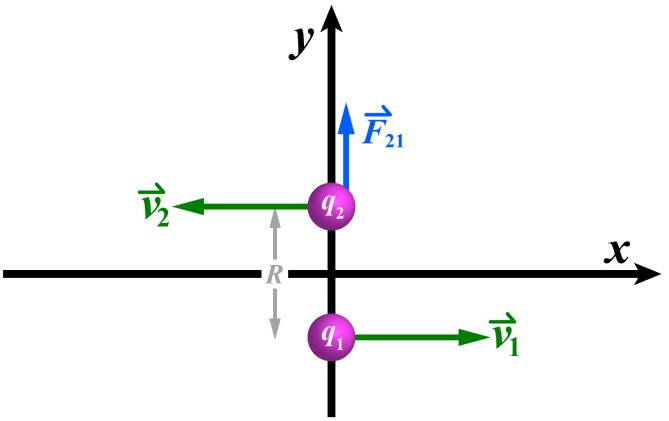


Figure 7.17: An IRF is shown in which two positive charges are moving in opposite directions parallel to the x-axis. On their closest approach they are separated by a distance R. The Lorentz force on q_2 due to q_1 is also shown.

where $k_M = k_E/c^2$, $u_1 = u$ is the speed of q_1, and $\vec{\phi} = \hat{z}$ in the xy-plane. Therefore, the Lorentz force on q_2 is given by Eq. 5.7.1 to be

$$\vec{F}_{21} = q_2\left(\vec{E}_1 + \vec{u}_2 \times \vec{B}_1\right)$$

where $q_2 = q$ is moving with a velocity of $\vec{u}_2 = -u\hat{x}$. Substituting in our electric and magnetic field equations, we get

$$\vec{F}_{21} = q\left(\gamma_1\, k_E \frac{q}{R^2}\hat{y} + \vec{u}_2 \times \left[\gamma_1\, \frac{k_E}{c^2}\frac{qu}{R^2}\hat{z}\right]\right)$$

$$\vec{F}_{21} = \gamma_1 k_E \frac{q^2}{R^2}\left(\hat{y} + \frac{u^2}{c^2}\left[-\hat{x} \times \hat{z}\right]\right).$$

Since $-\hat{x} \times \hat{z} = \hat{y}$ and $\beta \equiv u/c$,

$$\vec{F}_{21} = \gamma_1 k_E \frac{q^2}{R^2}\left(1 + \frac{u^2}{c^2}\right)\hat{y} = \gamma_1 k_E \frac{q^2}{R^2}\left(1 + \beta^2\right)\hat{y}$$

$$\vec{F}_{21} = \gamma_1\left(1 + \beta^2\right) k_E \frac{q^2}{R^2}\hat{y}.$$

- We know $\vec{r}'' = \vec{r}$ (to use label choices from Example 7.5.1) because there is no perceived length contraction. In the rest-frame of q_1, the E-field generated by q_1 at the location of q_2 is given exactly by Coulomb's law (defined by Eq. 5.2.5) to be

$$\vec{E}_1'' = k_E \frac{q_1}{r^3}\vec{r} = k_E \frac{q}{R^2}\hat{y}$$

because $q_1 = q$ and $\vec{r} = R\hat{y}$. There is no B-field because only moving charges generate B-fields (i.e. $\vec{B}_1'' = 0$). Therefore, the Lorentz force on q_2 is given by Eq. 5.7.1 to be

$$\vec{F}_{21}'' = q_2\left(\vec{E}_1'' + \vec{u}_2'' \times \vec{B}_1''\right)$$

where $q_2 = q$ is moving with a velocity of $\vec{u}_2''$. Substituting in our electric and magnetic field equations, we get

$$\vec{F}_{21}'' = q\left(k_E \frac{q}{R^2}\hat{y} + 0\right) = k_E \frac{q^2}{R^2}\hat{y}.$$

- Labeling the rest frame of q_2 as the single-primed frame, we also know $\vec{r}' = \vec{r}$ because there is no perceived length contraction. In this frame, the E-field generated by q_1 at the location of q_2 is given by Eq. 7.5.19 to be

$$\vec{E}_1' = \gamma_1' \, k_E \frac{q_1}{r^3}\vec{r} = \gamma_1' \, k_E \frac{q}{R^2}\hat{y}$$

because $\theta = 90°$, $q_1 = q$, $r = R$, and $\vec{r} = R\hat{y}$. To be clear on the notation used here, γ_1' is the gamma factor for q_1 from its own rest frame to the single-primed frame (i.e. $\gamma_1' \neq \gamma_1$). By the same logic, the B-field generated by q_1 at the location of q_2 is given by Eq. 7.5.21 to be

$$\vec{B}_1' = \gamma_1' \, k_M \frac{q u_1'}{R^2}\hat{\phi} = \gamma_1' \, \frac{k_E}{c^2}\frac{q u_1'}{R^2}\hat{z},$$

where $k_M = k_E/c^2$ and u_1' is the speed of q_1 in this frame. We know, from Eq. 7.3.3a, that

$$u_1' = \frac{u_1 - v}{1 - u_1 v/c^2} = \frac{u - (-u)}{1 - u\,(-u)\,/c^2} = \frac{2u}{1 + \beta^2},$$

where $\beta \equiv u/c$ and $v = -u$ is the relative velocity between the unprimed and single-primed frames. This makes the B-field

$$\vec{B}_1{}' = \gamma_1' \frac{k_E}{c^2}\frac{q}{R^2}\left(\frac{2u}{1+\beta^2}\right)\hat{z} = \left(\frac{2\gamma_1'}{1+\beta^2}\right)\frac{k_E}{c^2}\frac{qu}{R^2}\hat{z},$$

where γ_1' is related to u_1' by Eq. 7.4.4. Therefore, the Lorentz force on q_2 is given by Eq. 5.7.1 to be

$$\vec{F}_{21}' = q_2\left(\vec{E}_1{}' + \vec{u}_2' \times \vec{B}_1{}'\right)$$

where $q_2 = q$ is at rest (i.e. $\vec{u}_2' = 0$ resulting in no magnetic effect). Substituting in our electric and magnetic field equations, we get

$$\vec{F}_{21}' = q\left(\gamma_1' \, k_E \frac{q}{R^2}\hat{y} + 0\right) = \gamma_1' \, k_E \frac{q^2}{R^2}\hat{y}$$

- In the interest of comparing these three Lorentz forces, we'll need to know what each of the different gamma factors relate to each other through $\beta \equiv u/c$. Starting with the most complicated gamma factor, we get

$$\gamma_1' = \frac{1}{\sqrt{1-(u_1')^2/c^2}} = \frac{1}{\sqrt{1-\left(\frac{2\beta}{1+\beta^2}\right)^2}}$$

$$\gamma_1' = \frac{1+\beta^2}{\sqrt{(1+\beta^2)^2-(2\beta)^2}} = \frac{1+\beta^2}{\sqrt{1+2\beta^2+\beta^4-4\beta^2}}$$

$$\gamma_1' = \frac{1+\beta^2}{\sqrt{1-2\beta^2+\beta^4}} = \frac{1+\beta^2}{1-\beta^2} = \gamma_1^2\left(1+\beta^2\right).$$

That's almost the relativistic coefficient on the force in the unprimed frame! It only varies by a factor of γ_1. Well, it actually varies by a factor of γ_2, the gamma factor for q_2 between the unprimed frame and the rest frame of q_2 (the single-primed frame). This just so happens

to be equal to γ_1 in our example because the charges are traveling the same speed in the unprimed frame. In general, we can write

$$\vec{F}_{21} = \frac{\vec{F}'_{21}}{\gamma_2}, \tag{7.5.22}$$

which involves the rest frame of q_2. Also,

$$\vec{F}_{21} = \gamma_1 \left(1 + \beta^2\right) \vec{F}''_{21},$$

where neither frame in the transformation is the rest frame of the object (on which the force acts). This is much more complicated a transformations because the Lorentz force is a coordinate 3-force, so it doesn't transform between frames in the simple way that a 4-force would. Just for some perspective, if $u = 0.5c$ and $R = 10$ fm for two protons, then the three Lorentz forces have the values

$$\boxed{\begin{Bmatrix} F_{21} = 3.33 \text{ N} \\ F'_{21} = 3.85 \text{ N} \\ F''_{21} = 2.31 \text{ N} \end{Bmatrix}}$$

and it is clear that F'_{21} is the largest measured force.

- We can also discuss a few things more generally. Eq. 7.5.22 can be written as

$$\boxed{\vec{F}_{\perp} = \frac{\vec{F}_{p\perp}}{\gamma}}, \tag{7.5.23}$$

which is true of any force components such that those components are perpendicular to the motion of the object on which the force acts. The quantity $\vec{F}_p$ can be called the **proper force** (the maximum measurable force), which is measured in the rest frame of the object on which the force acts. As it turns out, the components parallel to the motion are measured the same in all frames.

Maxwell's Equations with Fields

Now that we have an EMF tensor, we can derive Maxwell's equations in terms of it. Unfortunately, with fields there will be two equations in the end (rather than just one like there was with potentials), so one could argue this wont be quite as elegant. In spacetime, we've grouped all the possible sources of the electromagnetic field into one quantity: the 4-current (given by Eq. 7.5.2). These sources show up in two of the four Maxwell's equations. Let's see if we can turn these two into one.

We'll start with Gauss's law since it results in a scalar and this will give us a simpler start. Eq. 5.4.9a states

$$\vec{\nabla} \bullet \vec{E} = \frac{\rho}{\epsilon_0} = \mu_0 c^2 \rho,$$

where we've used Eq. 5.5.4 to eliminate the fraction on the right side. If we divide through by $1/c$, then

$$\vec{\nabla} \bullet \left(\frac{\vec{E}}{c} \right) = \mu_0 \left(c\rho \right)$$

$$\nabla_x \left(\frac{E_x}{c} \right) + \nabla_y \left(\frac{E_y}{c} \right) + \nabla_z \left(\frac{E_z}{c} \right) = \mu_0 \left(c\rho \right).$$

On the left, we have three spatial terms from a scalar product, which in spacetime should also involve time. Since $\mathcal{F}^{00} = 0$, we can perform something I like to call *voodoo math* (with a little foresight; we can add zeros, multiply by ones, add and subtract constants, etc. to simplify a mathematical expression). Using this and Eq. 7.5.2, we can write Gauss's law as

$$\nabla_\delta \mathcal{F}^{0\delta} = \mu_0 J^0, \tag{7.5.24}$$

where δ is a summation index and $\mathcal{F}^{0\delta}$ represents a component in the zeroth row of the contravariant EMF tensor (given by Eq. 7.5.12).

The other of Maxwell's equations involving sources of fields is Ampére's law (given by Eq. 5.4.9d), which states

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{J} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t}$$

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{J} + \frac{1}{c^2} \frac{\partial \vec{E}}{\partial t},$$

where we've used Eq. 5.5.4 to simplify a term on the right side. With a little manipulation, we get

$$\vec{\nabla} \times \vec{B} - \frac{1}{c^2} \frac{\partial \vec{E}}{\partial t} = \mu_0 \vec{J}$$

$$\vec{\nabla} \times \vec{B} - \frac{1}{c} \frac{\partial}{\partial t} \left(\frac{E_x}{c} \right) = \mu_0 \vec{J},$$

which gets all the field information on the left side. This equation is actually three equations, one for each spatial component of the vectors. If we intend to write this in index notation, then we need to have them all separate leaving us with

$$\left\{ \begin{array}{l} \left(\vec{\nabla} \times \vec{B} \right)_x - \frac{1}{c} \frac{\partial}{\partial t} \left(\frac{E_x}{c} \right) = \mu_0 J_x \\ \left(\vec{\nabla} \times \vec{B} \right)_y - \frac{1}{c} \frac{\partial}{\partial t} \left(\frac{E_y}{c} \right) = \mu_0 J_y \\ \left(\vec{\nabla} \times \vec{B} \right)_z - \frac{1}{c} \frac{\partial}{\partial t} \left(\frac{E_z}{c} \right) = \mu_0 J_z \end{array} \right\}$$

noting that $B_i = B^i$ in Cartesian 3-space due to Eq. 6.4.5. Using the definitions of the cross product (Eq. 2.2.4) and the covariant derivative (Eq. 7.5.1), we get

$$\left\{ \begin{array}{l} (\nabla_y B_z - \nabla_z B_y) - \nabla_t \left(E_x/c \right) = \mu_0 J_x \\ (\nabla_z B_x - \nabla_x B_z) - \nabla_t \left(E_y/c \right) = \mu_0 J_y \\ (\nabla_x B_y - \nabla_y B_x) - \nabla_t \left(E_z/c \right) = \mu_0 J_z \end{array} \right\}$$

$$\left\{ \begin{array}{l} \nabla_y B_z + \nabla_z \left(-B_y \right) + \nabla_t \left(-E_x/c \right) = \mu_0 J_x \\ \nabla_z B_x + \nabla_x \left(-B_z \right) + \nabla_t \left(-E_y/c \right) = \mu_0 J_y \\ \nabla_x B_y + \nabla_y \left(-B_x \right) + \nabla_t \left(-E_z/c \right) = \mu_0 J_z \end{array} \right\}.$$

Based on the form of the contravariant EMF tensor (given by Eq. 7.5.12), we can write this as

$$\left\{ \begin{array}{l} \nabla_2 \mathcal{F}^{12} + \nabla_3 \mathcal{F}^{13} + \nabla_0 \mathcal{F}^{10} = \mu_0 J^1 \\ \nabla_3 \mathcal{F}^{23} + \nabla_1 \mathcal{F}^{21} + \nabla_0 \mathcal{F}^{20} = \mu_0 J^2 \\ \nabla_1 \mathcal{F}^{31} + \nabla_2 \mathcal{F}^{32} + \nabla_0 \mathcal{F}^{10} = \mu_0 J^3 \end{array} \right\}.$$

Since $\mathcal{F}^{11} = \mathcal{F}^{22} = \mathcal{F}^{33} = 0$ (the missing term from each of the summations), we can perform something I like to call *voodoo math* (with a little foresight; we can add zeros, multiply by ones, add and subtract constants, etc. to simplify a mathematical expression). Using this and Eq. 7.5.2, we can write Ampére's law as

$$\left\{\begin{array}{l} \nabla_\delta \mathcal{F}^{1\delta} = \mu_0 J^1 \\ \nabla_\delta \mathcal{F}^{2\delta} = \mu_0 J^2 \\ \nabla_\delta \mathcal{F}^{3\delta} = \mu_0 J^3 \end{array}\right\}. \tag{7.5.25}$$

The components given in Eq. 7.5.24 and Eq. Set 7.5.25 have an identical form, so we can combine them into one equation using index notation. This results in

$$\nabla_\delta \mathcal{F}^{\alpha\delta} = \mu_0 J^\alpha, \tag{7.5.26}$$

where δ is a summation index and α is a free index. I would argue this is elegant in its simplicity even if it doesn't represent a complete description of electrodynamics. As was mentioned earlier, there are two other Maxwell's equations: the ones without sources in them. These correspond to Faraday's law and Gauss's law for magnetism.

Starting again with the scalar product for simplicity, Gauss's law for magnetism (given by Eq. 5.4.9b) states

$$\vec{\nabla} \bullet \vec{B} = \nabla_x B_x + \nabla_y B_y + \nabla_z B_z = 0.$$

We can use the covariant EMF tensor (given by Eq. 7.5.13) to write this as

$$\nabla_1 \mathcal{F}_{23} + \nabla_2 \mathcal{F}_{31} + \nabla_3 \mathcal{F}_{12} = 0; \tag{7.5.27}$$

where 123, 231, and 321 are the even permutations of the indices. This one was probably the easiest so far.

Faraday's law is a vector equation and, therefore, has three components like Ampére's law. Eq. 5.4.9c states

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

$$\vec{\nabla} \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = 0,$$

where we've manipulated a bit to get all the field information on the left side. We can now multiply through by $1/c$ to achieve the right units and the result is

$$\vec{\nabla} \times \left(\frac{\vec{E}}{c} \right) + \frac{1}{c} \frac{\partial \vec{B}}{\partial t} = 0.$$

If we intend to write this in index notation, then we need to have all the components separate leaving us with

$$\left\{ \begin{array}{l} \left[\vec{\nabla} \times \left(\frac{\vec{E}}{c} \right) \right]_x + \frac{1}{c} \frac{\partial B_x}{\partial t} = 0 \\ \left[\vec{\nabla} \times \left(\frac{\vec{E}}{c} \right) \right]_y + \frac{1}{c} \frac{\partial B_y}{\partial t} = 0 \\ \left[\vec{\nabla} \times \left(\frac{\vec{E}}{c} \right) \right]_z + \frac{1}{c} \frac{\partial B_z}{\partial t} = 0 \end{array} \right\},$$

noting that $B_i = B^i$ in Cartesian 3-space due to Eq. 6.4.5. Using the definitions of the cross product (Eq. 2.2.4) and the covariant derivative (Eq. 7.5.1), we get

$$\left\{ \begin{array}{l} \nabla_y \left(E_z/c \right) - \nabla_z \left(E_y/c \right) + \nabla_t B_x = 0 \\ \nabla_z \left(E_x/c \right) - \nabla_x \left(E_z/c \right) + \nabla_t B_y = 0 \\ \nabla_x \left(E_y/c \right) - \nabla_y \left(E_x/c \right) + \nabla_t B_z = 0 \end{array} \right\}$$

$$\left\{ \begin{array}{l} \nabla_y \left(E_z/c \right) + \nabla_z \left(-E_y/c \right) + \nabla_t B_x = 0 \\ \nabla_z \left(E_x/c \right) + \nabla_x \left(-E_z/c \right) + \nabla_t B_y = 0 \\ \nabla_x \left(E_y/c \right) + \nabla_y \left(-E_x/c \right) + \nabla_t B_z = 0 \end{array} \right\}$$

Based on the form of the covariant EMF tensor (given by Eq. 7.5.13), we can write this as

$$\left\{ \begin{array}{l} \nabla_2 \mathcal{F}_{30} + \nabla_3 \mathcal{F}_{02} + \nabla_0 \mathcal{F}_{23} = 0 \\ \nabla_3 \mathcal{F}_{10} + \nabla_1 \mathcal{F}_{03} + \nabla_0 \mathcal{F}_{31} = 0 \\ \nabla_1 \mathcal{F}_{20} + \nabla_2 \mathcal{F}_{01} + \nabla_0 \mathcal{F}_{12} = 0 \end{array} \right\} \tag{7.5.28}$$

where again we have even permutations of the indices in each component equation.

The components given in Eq. 7.5.27 and Eq. Set 7.5.28 have an identical form, so we can combine them into one equation using index notation. This results in

$$\nabla_\alpha \mathcal{F}_{\nu\delta} + \nabla_\nu \mathcal{F}_{\delta\alpha} + \nabla_\delta \mathcal{F}_{\alpha\nu} = 0\,; \tag{7.5.29}$$

where α, ν, and δ are all free indices. This complete's our derivation of Maxwell's equations, but does it give us a complete description of of electrodynamics? The answer is a resounding "No." Just as in Section 5.4, we need to know how charges will respond to these fields and that requires the Lorentz force.

Lorentz Four-Force

In vector notation, the Lorentz 3-force is given by Eq. 5.7.1 as

$$\vec{F} = q\left(\vec{E} + \vec{u} \times \vec{B}\right)$$

where $\vec{u}$ is the velocity of q. We also define the parenthetical quantity as the electromagnetic field. In this section however, we write the electromagnetic field as a rank-2 tensor given by Eq. 7.5.12, so we'll need to rewrite the Lorentz force as a 4-vector in index notation. We'll call this the **Lorentz 4-Force**.

Judging from its appearance, it also involves charge and velocity. The quantities are multiplied, so it stands to reason that they will also multiply in index notation. Let's try

$$F^\delta = q u_\alpha \mathcal{F}^{\delta\alpha}\,, \tag{7.5.30}$$

where ν is a summation index and δ is a free index. The quantity u_ν is the covariant 4-velocity given by

$$u_\alpha = g_{\alpha\delta} u^\delta \longrightarrow (-\gamma c, \gamma \vec{u})\,,$$

which only differs from the contravariant 4-velocity by the negative sign on the time component. Checking this 4-vector's spatial components, we get

$$\left\{\begin{array}{l} F^1 = q\left(u_0\mathcal{F}^{10} + u_1\mathcal{F}^{11} + u_2\mathcal{F}^{12} + u_3\mathcal{F}^{13}\right) \\ F^2 = q\left(u_0\mathcal{F}^{20} + u_1\mathcal{F}^{21} + u_2\mathcal{F}^{22} + u_3\mathcal{F}^{23}\right) \\ F^3 = q\left(u_0\mathcal{F}^{30} + u_1\mathcal{F}^{31} + u_2\mathcal{F}^{32} + u_3\mathcal{F}^{33}\right) \end{array}\right\}$$

or more simply

$$\left\{\begin{array}{l} F^1 = q\left(u_0\mathcal{F}^{10} + u_2\mathcal{F}^{12} + u_3\mathcal{F}^{13}\right) \\ F^2 = q\left(u_0\mathcal{F}^{20} + u_1\mathcal{F}^{21} + u_3\mathcal{F}^{23}\right) \\ F^3 = q\left(u_0\mathcal{F}^{30} + u_1\mathcal{F}^{31} + u_2\mathcal{F}^{32}\right) \end{array}\right\},$$

where we've made $\mathcal{F}^{11} = \mathcal{F}^{22} = \mathcal{F}^{33} = 0$. With the components of the contravariant EMF tensor and the covariant 4-velocity, these components become

$$\left\{\begin{array}{l} F^1 = q\left[-\gamma c\left(-E_x/c\right) + \gamma u_y\left(B_z\right) + \gamma u_z\left(-B_y\right)\right] \\ F^2 = q\left[-\gamma c\left(-E_y/c\right) + \gamma u_x\left(-B_z\right) + \gamma u_z\left(B_x\right)\right] \\ F^3 = q\left[-\gamma c\left(-E_z/c\right) + \gamma u_x\left(B_y\right) + \gamma u_y\left(-B_x\right)\right] \end{array}\right\}$$

$$\left\{\begin{array}{l} F^1 = \gamma q\left[E_x + u_y B_z - u_z B_y\right] \\ F^2 = \gamma q\left[E_y - u_x B_z + u_z B_x\right] \\ F^3 = \gamma q\left[E_z + u_x B_y - u_y B_x\right] \end{array}\right\}.$$

By the definition of the cross product (Eq. 2.2.4), this becomes

$$\boxed{\left\{\begin{array}{l} F^1 = \gamma q\left(E_x + \left[\vec{u}\times\vec{B}\right]_x\right) \\ F^2 = \gamma q\left(E_y + \left[\vec{u}\times\vec{B}\right]_y\right) \\ F^3 = \gamma q\left(E_z + \left[\vec{u}\times\vec{B}\right]_z\right) \end{array}\right\}},$$

which is almost exactly the components of the Lorentz 3-force. The extra factor of γ is consistent with Eq. 7.4.28 because the original Lorentz 3-force is a coordinate force (i.e. involved coordinate time, not proper time).

This is only three components. What about the time component of the Lorentz 4-force? By the same methods as above, it is

$$F^0 = q\left(u_0\mathcal{F}^{00} + u_1\mathcal{F}^{01} + u_2\mathcal{F}^{02} + u_3\mathcal{F}^{03}\right)$$

$$F^0 = q\left(u_1\mathcal{F}^{01} + u_2\mathcal{F}^{02} + u_3\mathcal{F}^{03}\right)$$

$$F^0 = q\left[\gamma u_x\left(\frac{E_x}{c}\right) + \gamma u_y\left(\frac{E_y}{c}\right) + \gamma u_z\left(\frac{E_z}{c}\right)\right]$$

$$F^0 = \gamma \frac{q}{c} \left(u_x E_x + u_y E_y + u_z E_z \right) = \gamma \frac{q}{c} \left(\vec{u} \bullet \vec{E} \right),$$

where we've used the definition of the dot product (Eq. 2.2.2). There still may be some confusion as to what this is, but if we bring in the q, then

$$F^0 = \frac{\gamma}{c} \left(\vec{u} \bullet q\vec{E} \right) = \frac{\gamma}{c} \left(\vec{u} \bullet \vec{F}_E \right). \tag{7.5.31}$$

We know from classical physics that $P = \vec{u} \bullet \vec{F}$. The parenthetical quantity is just the **coordinate electrical power**! The factor of γ/c is consistent with Eq. 7.4.28. It also makes sense that the magnetic field is not involved in power because it never does work:

$$P = \vec{u} \bullet \vec{F}_B = \vec{u} \bullet \left(q\vec{u} \times \vec{B} \right) = q\vec{u} \bullet \left(\vec{u} \times \vec{B} \right) = 0,$$

which is true for any $\vec{u}$ or $\vec{B}$. A more clear way to look at Eq. 7.5.31 than just calling it electrical power is to say it's the rate at which energy is added to the charge q by the electric field.

Example 7.5.3

Back in Example 7.5.2, we had two equal positive charges moving in opposite directions and found the Lorentz 3-force one due to the other in three different frames. Find the Lorentz 4-force on the same charge in those same three frames.

- To keep this short, we'll be using a lot from Example 7.5.2 (i.e. reference said example if you feel like there are gaps in this one). We've already gone through a little work with the Lorentz 4-force, so we'll start from

$$\left\{ \begin{array}{l} F^0 = F^t = \gamma \dfrac{q}{c} \left(u_x E_x + u_y E_y + u_z E_z \right) \\ \\ F^1 = F^x = \gamma q \left(E_x + u_y B_z - u_z B_y \right) \\ F^2 = F^y = \gamma q \left(E_y - u_x B_z + u_z B_x \right) \\ F^3 = F^z = \gamma q \left(E_z + u_x B_y - u_y B_x \right) \end{array} \right\}.$$

$$\left\{ \begin{array}{l} F^t = \gamma_2 \dfrac{q_2}{c} \left(u_{2x} E_{1x} + u_{2y} E_{1y} + u_{2z} E_{1z} \right) \\ \\ F^x = \gamma_2 q_2 \left(E_{1x} + u_{2y} B_{1z} - u_{2z} B_{1y} \right) \\ F^y = \gamma_2 q_2 \left(E_{1y} - u_{2x} B_{1z} + u_{2z} B_{1x} \right) \\ F^z = \gamma_2 q_2 \left(E_{1z} + u_{2x} B_{1y} - u_{2y} B_{1x} \right) \end{array} \right\},$$

where subscripts of 1 correspond to q_1 and subscripts of 2 correspond to q_2. Since we know from Example 7.5.2 that $E_{1x} = E_{1z} = B_{1x} = B_{1y} = 0$ and $u_{2y} = u_{2z} = 0$, then we get

$$F^y = \gamma_2 \, q \, (E_1 - u_2 B_1)$$

where $q_2 = q$ is invariant and the rest of the terms are zero as we'd expect. The fields E_1 and B_1 are given by Eqs. 7.5.19 and 7.5.21, respectively.

- In the IRF in which the two charges are traveling the same speed (i.e. the unprimed frame), we know $u_1 = u$ and $u_2 = -u$. We also know

$$\left\{ \begin{array}{l} E_1 = \gamma_1 \, k_E \dfrac{q}{R^2} \\ \\ B_1 = \gamma_1 \, \dfrac{k_E}{c^2} \dfrac{qu}{R^2} \end{array} \right\},$$

so

$$F^y = \gamma_2 \, q \left(\gamma_1 \, k_E \frac{q}{R^2} + u \gamma_1 \, \frac{k_E}{c^2} \frac{qu}{R^2} \right)$$

$$F^y = \gamma_2 \gamma_1 \left(1 + \frac{u^2}{c^2} \right) k_E \frac{q^2}{R^2} = \gamma_2 \gamma_1 \left(1 + \beta^2 \right) k_E \frac{q^2}{R^2},$$

which is exactly what we got in Example 7.5.2 with the extra factor of γ_2 we expect from Eq. 7.4.28.

- In rest frame of q_1 (i.e. the double-primed frame), we know $u_1'' = 0 \Rightarrow \gamma_1'' = 1$ and, from Eq. 7.3.3a, that

$$u_2'' = \frac{u_2 - v}{1 - u_2 v / c^2} = \frac{(-u) - u}{1 - (-u) \, u / c^2} = \frac{-2u}{1 + \beta^2},$$

where $\beta \equiv u/c$ and $v = u$ is the relative velocity between the unprimed and double-primed frames. We also know

$$\left\{ \begin{array}{c} E_1'' = k_E \dfrac{q}{R^2} \\ \\ B_1'' = 0 \end{array} \right\},$$

so

$$F''^{\,y} = \gamma_2'' \, q \left(k_E \frac{q}{R^2} + 0 \right) = \gamma_2'' \, k_E \frac{q^2}{R^2}$$

where

$$\gamma_2'' = \frac{1}{\sqrt{1 - u_2''/c^2}},$$

which is exactly what we got in Example 7.5.2 with the extra factor of γ_2'' we expect from Eq. 7.4.28.

- In rest frame of q_2 (i.e. the single-primed frame), we know $u_2' = 0 \Rightarrow \gamma_2' = 1$ and, from Eq. 7.3.3a, that

$$u_1' = \frac{u_1 - v}{1 - u_1 v/c^2} = \frac{u - (-u)}{1 - u\,(-u)\,/c^2} = \frac{2u}{1 + \beta^2},$$

where $\beta \equiv u/c$ and $v = -u$ is the relative velocity between the unprimed and single-primed frames. We also know

$$\left\{ \begin{array}{l} E_1' = \gamma_1' \, k_E \dfrac{q}{R^2} \\ \\ B_1' = \gamma_1' \, \dfrac{k_E}{c^2} \dfrac{q u_1}{R^2} \end{array} \right\},$$

so

$$F'^{\,y} = q \left(\gamma_1' \, k_E \frac{q}{R^2} + 0 \right) = \gamma_1' \, k_E \frac{q^2}{R^2}$$

where

$$\gamma_1' = \frac{1}{\sqrt{1 - u_1'/c^2}},$$

which is exactly what we got in Example 7.5.2 with no extra factor because $\gamma_2'' = 1$.

- Ok, so we got what we expected given our results in Example 7.5.2. We also plugged in some numbers: $u = 0.5c$ and $R = 10$ fm for two protons. This results in

$$\boxed{\left\{ \begin{array}{lcl} F^{\delta} & \longrightarrow & (0,\ 3.85\ \text{N}\ \hat{y}) \\ F'^{\,\delta} & \longrightarrow & (0,\ 3.85\ \text{N}\ \hat{y}) \\ F''^{\,\delta} & \longrightarrow & (0,\ 3.85\ \text{N}\ \hat{y}) \end{array} \right\}}$$

using the $(\text{time}, \overrightarrow{\text{space}})$ shorthand. It would *appear* as though the Lorentz 4-force is invariant. However, they are only the same because the electric field is orthogonal to the motion of both charges. We can show this kind of transformation in matrix notation as

$$\begin{bmatrix} 0 \\ 0 \\ F'^{y} \\ F'^{z} \end{bmatrix} = \begin{bmatrix} \gamma_T & \pm\gamma_T\beta_T & 0 & 0 \\ \pm\gamma_T\beta_T & \gamma_T & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ F^{y} \\ F^{y} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ F^{y} \\ F^{y} \end{bmatrix}$$

where the $\pm$ indicates the transformation can occur in either direction. If there is a component along the direction of motion (i.e. $F^x \neq 0$), then we also know there is a time component (i.e. $F^t \neq 0$) by Eq. 7.5.31 and the transformation will not leave the Lorentz 4-force invariant.

- It should also be noted that the time component will not remain zero as time passes because q_2 (and q_1 for that matter) will gradually gain a u_y component. Furthermore, the moment each of these charge experiences a 4-force, their rest frames are no longer IRFs. That means this work is only valid if these charges were held in the same rest frame by some outside force *then* instantly shifted into their different frames at beginning of the example and *even then* it still only applies to that moment. We can only transform between IRFs and the only frame that remains an IRF is the one in between the two rest frames (i.e. the unprimed frame). The chance of this scenario occurring in the real universe is highly unlikely.

7.6 Worldines

Everything we've done so far in this chapter has been objects traveling along time-like world lines. This isn't a horrible place to start an understanding since almost everything we interact with in our everyday life travels these. However, as we've mentioned before, not everything does. Addressing these circumstances requires us to step outside our comfort zone and look at the universe as objectively as possible.

Null World Lines

Particles that travel at speeds *exactly* equal to c follow **null world lines** meaning the spacetime separation between events they interact with is zero. We mentioned in Example 7.4.1 that we weren't prepared to deal with particles (or objects) traveling at c, but there was no real explanation as to why. First, consider a particle (of unknown mass) and assume it is traveling at $u = c$ (or $\beta = 1$) in the x-direction in some IRF. It's coordinate 3-momentum in that same IRF would be

$$\vec{p} = m_p \vec{u} = m_p u \hat{x} = m_p c \hat{x},$$

which easily has a finite value. No problems, right? However, it's relativistic 3-momentum is given by

$$\vec{p}_{\text{rel}} = \gamma \vec{p} = \frac{m_p c \hat{x}}{\sqrt{1 - \beta^2}},$$

where we've already said $\beta = 1$.

Here in lies our problem. If $\beta = 1$, then the denominator is zero and $\vec{p}_{\text{rel}} = \infty$. Since nothing can *actually* have an infinite value in the real physical universe, we can conclude that $\beta \neq 1$ (a proof by contradiction). Therefore, we can approach speeds of c, but can never actually accelerate to *exactly* c. The muon-antineutrino in Example 7.4.1 got pretty close, but it still didn't reach what we consider to be the universal speed limit.

You might be thinking "What?! Photons travel at the speed of light!" and indeed they do. How they do it is the better question. Photons have a zero rest mass (i.e. $m_p = 0$) resulting in a coordinate 3-momentum of

$$\vec{p} = m_p u \vec{x} = (0)\, c \hat{x} = 0$$

and a relativistic 3-momentum of

$$\vec{p}_{\text{rel}} = \frac{m_p c \hat{x}}{\sqrt{1 - \beta^2}} = \frac{0}{0},$$

which is called an **indeterminate form** in mathematics. The abstract details of the indeterminate form are unimportant. The important thing is, however indeterminate it might be, it has a finite result. Therefore, the photon (and any other particle with $m_p = 0$) does not violate our system of mathematics.

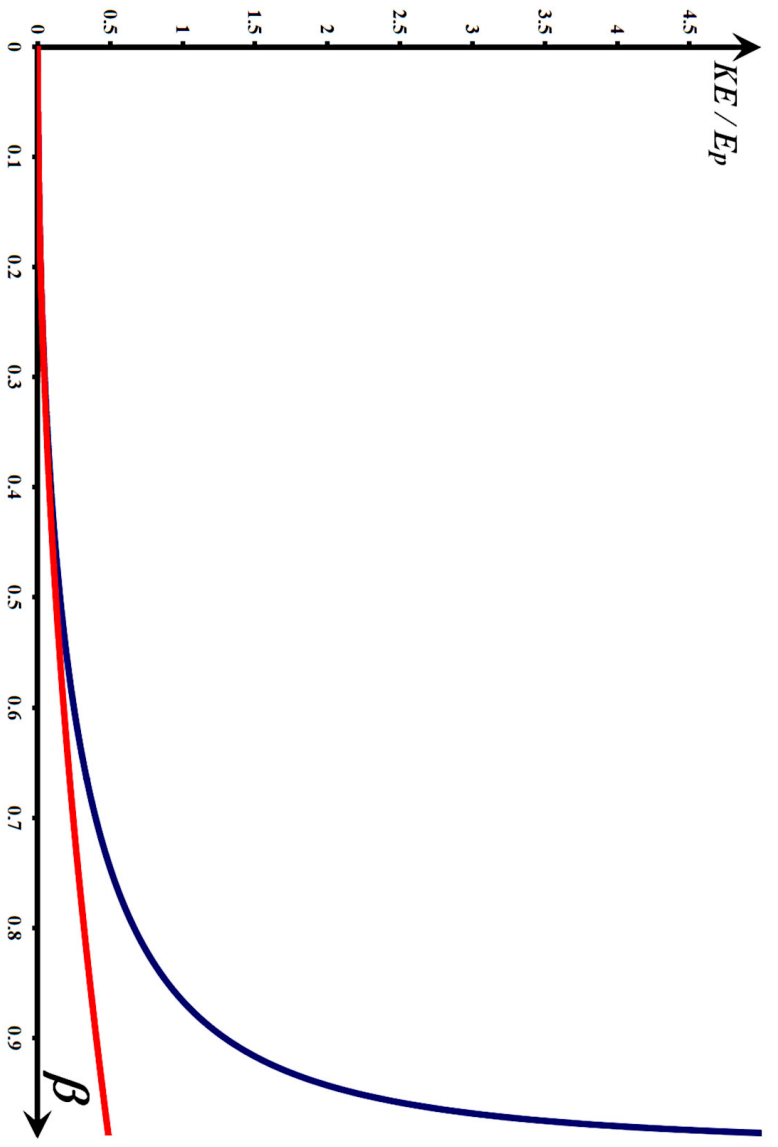


Figure 7.18: This is a graph of kinetic energy (KE/E_p) vs. velocity (β) scaled so that both axes are unitless. The blue curve is the relativistic kinetic energy, which goes to infinity at $\beta = 1$ indicating that it requires an infinite amount of energy to accelerate to $v = c$. The red curve is the classical version, which visibly begins to deviate from the more accurate relativistic version at about $\beta = 0.4$.

Ok, so massive particles *always* travel along time-like world lines and zero rest mass particles *always* travel along null world lines. What does it means for two events to have a null separation? According to the line element (Eq. 7.2.1), this separation would be

$$0 = (\Delta s)^2 = -c^2 (\Delta t)^2 + (\Delta x)^2$$

in the IRF of this discussion. This corresponds to

$$c^2 (\Delta t)^2 = (\Delta x)^2 \quad \Rightarrow c\Delta t = \Delta x \tag{7.6.1}$$

implying that the time and space components of 4-vectors along null world lines will have the same value. We saw this occur approximately with the 4-momentum of muon-antineutrino in Example 7.4.1. If we take the scalar product of that 4-momentum with itself (using Eq. 7.4.15), then we'd get

$$p_\delta p^\delta = -\left(29.8\,\frac{\text{MeV}}{c}\right)^2 + \left(-29.8\,\frac{\text{MeV}}{c}\right)^2 \hat{x} \bullet \hat{x} \approx 0,$$

which makes sense considering neutrinos are nearly massless. You could also argue this in general using Eq. 7.4.24, resulting in

$$p_\delta p^\delta = -m_p^2 c^2 \approx 0$$

for nearly massless particles.

However, this zero result for the scalar product is true of *all* 4-vectors for massless particles due to Eq. 7.6.1, which is why we call them **null vectors**. We can get another useful result using Eq. 7.4.25. By substituting $m_p = 0$, we get

$$E_{\text{rel}}^2 = p_{\text{rel}}^2 c^2 \quad \Rightarrow E_{\text{rel}} = p_{\text{rel}} c$$

$$\boxed{p_{\text{rel}} = \frac{E_{\text{rel}}}{c}} \tag{7.6.2}$$

for all zero rest mass particles (note: $E_{\text{rel}} = hf_{\text{rel}}$ for a photon). We can also use this to write the 4-momentum as

$$p^\delta \longrightarrow \left(\frac{E_{\text{rel}}}{c},\ \frac{E_{\text{rel}}}{c}\hat{u}\right) \tag{7.6.3}$$

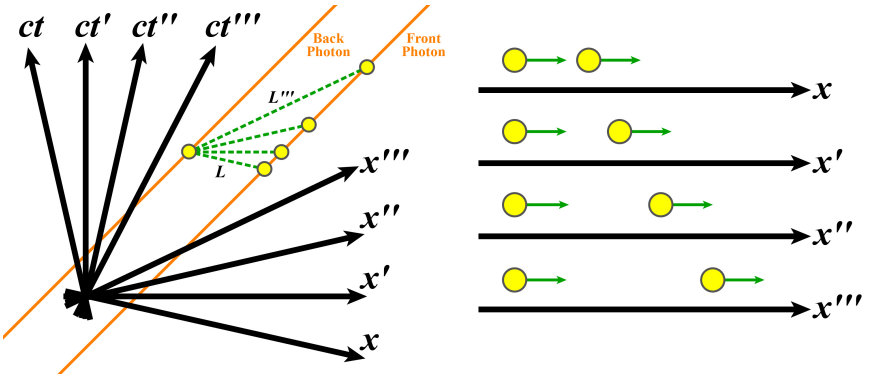


Figure 7.19: On the left is a spacetime diagram that includes four different IRFs observing the motion of two photons along the x-axis. The spacing between these photons is defined as the spacetime separation connecting two simultaneous events. On the right, you can see how the spacing between the photons gets larger as you approach the rest frame of the photons. Since there is no maximum value for length, proper length does not exist.

using the $(\text{time}, \overrightarrow{\text{space}})$ shorthand, where $\hat{u}$ is the direction of motion.

Now, let's shift perspective to the rest frame of this zero rest mass particle. A particle traveling at c even having a rest frame is a strange concept because the speed of light is a spacetime invariant, but let's consider it anyway. According to the line element (Eq. 7.2.1), the separation between two events would be

$$0 = (\Delta s)^2 = -c^2 (\Delta \tau)^2 \quad \Rightarrow \Delta \tau = 0$$

meaning no time passes *at all* for a zero rest mass particle. This is still consistent with time dilation because Eq. 7.2.11 says

$$\Delta t = \gamma \, \Delta \tau = \frac{\Delta \tau}{\sqrt{1 - \beta^2}} = \frac{0}{0},$$

which again is indeterminate resulting in a finite value for Δt. This also implies the entire concept of proper length is meaningless. Two photons can be spaced by a finite distance in every IRF except the rest frame of the photons (See Figure 7.19).

Having zero proper time poses a much larger problem for us. In Section 7.4, we defined all the 4-vectors as derivatives with respect to proper time, $d\tau$. A differential must be a very small number, but *not* zero, by definition. For massive particles, we were essentially using τ as a parameter (or independent

variable) to relate the coordinates (ct, x, y, z). We could have chosen anything really, but τ was convenient because it made sense dimensionally and gave us the relativistic form of Newton's first law in Eq. 7.4.29.

For zero rest mass particles, we'll have to resort to choosing something else. Choosing this new parameter carefully, we can get

$$u^\delta = \frac{dx^\delta}{d\Omega} = \text{constant} \quad \left(\text{if } F^\delta = 0\right), \tag{7.6.4}$$

where Ω is called an **affine parameter**. An affine parameter is simply a parameter which keeps the form of Newton's first law, so it isn't all that special but it is useful. There is no single value of Ω that will make it affine, so it's a bit more abstract than τ. With this in mind, definitions for 4-acceleration and 4-force follow as

$$a^\delta = \frac{du^\delta}{d\Omega} \text{ and } F^\delta = \frac{dp^\delta}{d\Omega}$$

If you're not feeling comfortable with there being a F^δ on something like a photon, then recall **Compton scattering**. When the photon scatters off a massive particle like an electron, there is most definitely a change in its 4-momentum. If a photon's frequency changes, then by $E = hf$ its energy will also change and energy is a part of 4-momentum (See Eq. 7.6.3).

Space-Like World Lines

We've now tackled *massive* particles on time-like world lines and *massless* particles on null (or light-like) world lines. That leaves one option remaining: particles on **space-like world lines**. We often call these particles **tachyons** and, since their introduction, they have become the basis of many science-fiction ideas. Before we get started in analyzing tachyons, I'd like to emphasize that they are *pure fantasy* at this point because they have *never* been experimentally detected. It is a common (and logically sound) policy in science to assume the non-existence of something prior to its discovery, but also be prepared to accept its existence upon discovery. It is the goal the following work to prepare you for the possibility of the existence of the tachyon.

Let's consider a particle (of unknown mass) and assume it is traveling at $u > c$ (or $\beta > 1$) in the x-direction in some IRF. It's coordinate 3-momentum

in that same IRF would be

$$\vec{p} = m_p \vec{u} = m_p u \hat{x}$$

which easily has a ordinary value. No problems, right? However, it's relativistic 3-momentum is given by

$$\vec{p}_{\text{rel}} = \gamma \vec{p} = \frac{m_p u \hat{x}}{\sqrt{1-\beta^2}},$$

and here in lies our problem. If $\beta > 1$, then the quantity under the square is negative and $\vec{p}_{\text{rel}}$ is imaginary (as well as many other quantities involving γ). In an attempt to avoid this problem, we can assume rest mass is imaginary for tachyons (i.e. $m_p = i z_p$), which gives us

$$\vec{p}_{\text{rel}} = \frac{i z_p u \hat{x}}{i\sqrt{\beta^2-1}} = \frac{z_p u \hat{x}}{\sqrt{\beta^2-1}}, \tag{7.6.5}$$

which is once again real. Furthermore, we can say

$$E_{\text{rel}} = \frac{i z_p c^2}{i\sqrt{\beta^2-1}} = \frac{z_p c^2}{\sqrt{\beta^2-1}} \tag{7.6.6}$$

and Eq. 7.4.25 becomes

$$E_{\text{rel}}^2 = (i z_p)^2 c^4 + p_{\text{rel}}^2 c^2 = -z_p^2 c^4 + p_{\text{rel}}^2 c^2$$

$$E_{\text{rel}}^2 + z_p^2 c^4 = p_{\text{rel}}^2 c^2. \tag{7.6.7}$$

It's clear now that, at least mathematically, special relativity doesn't discount the existence of such particles, but what kinds of consequences would their existence present?

Example 7.6.1

Consider two experimenters, Joe and Ashley, moving at a constant relative velocity $v = 0.8c$ with respect to each other. Joe sends Ashley a message saying "What's up?" using a radio wave. Upon receiving Joe's message, Ashley replies with "Nothin' much." using a radio wave. Assuming they've both accounted for the Doppler effect of light, they can both receive and send

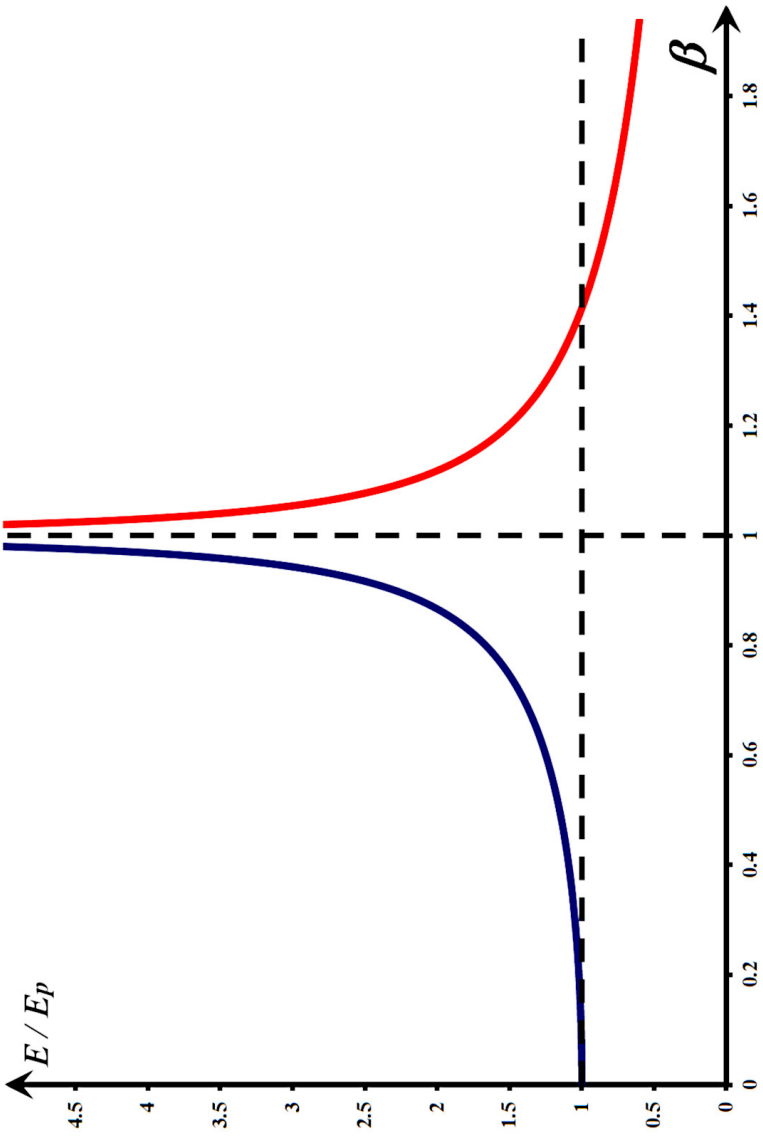


Figure 7.20: This is a graph of energy (E/E_p) vs. velocity (β) scaled so that both axes are unitless. A horizontal dashed line indicates the rest energy of each particle. The blue curve is the relativistic total energy of a massive real particle, which goes to infinity at $\beta = 1$ as it did in Figure 7.18. The red curve is the relativistic total energy of an imaginary tachyon, which goes to infinity at $\beta = 1$ showing that energy increases as speed decreases. Another interesting result is the tachyon's total energy can be less than its rest energy if it goes fast enough.

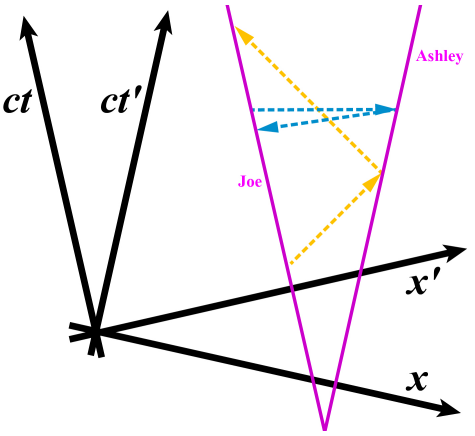


Figure 7.21: This is a spacetime diagram of two experimenters, Joe and Ashley, sending signals to each other. The orange dashed arrows represent radio waves (photons) being sent between them. The blue dashed arrows represent tachyons being sent between them. The tachyons travel into the future in one IRF, but the past in another IRF.

a signal at $u = c$ (which they'll both measure the same since it's a spacetime invariant).

Now consider the same two experimenters still moving at a constant relative velocity $v = 0.8c$ with respect to each other. Joe sends Ashley a message saying "What's up?" using tachyons traveling at $u = 5c$ (yes, I said five). They have both agreed that, upon receiving the message from Joe, Ashley will reply with "Don't send your message." using tachyons of equal speed (measured relative to her frame, of course). According to the spacetime diagram in Figure 7.21, the reply Ashley sends will travel forward in time in her IRF, but back in time in Joe's. Joe will receive this reply before he sends his original message and we have a **causality** problem.

This might be surpassing the limitation of the spacetime diagram, so let's do the problem with Lorentz transformations instead. According to Eq. Set 7.3.2,

$$\Delta t' = \gamma_T \left(\Delta t - \frac{v \Delta x}{c^2} \right)$$

$$c\Delta t' = \gamma_T \left(c\Delta t - \beta_T \Delta x \right),$$

where $\beta_T = v/c = 0.8$. For the original signal Joe sent,

$$\beta = \frac{u}{c} = \frac{\Delta x/\Delta t}{c} = \frac{\Delta x}{c\Delta t} \quad \Rightarrow \Delta x = \beta c \Delta t,$$

where $\beta = 5$. Now the time transformation is

$$c\Delta t' = \gamma_T \left(c\Delta t - \beta_T \beta \, c\Delta t\right) = \gamma_T \left(1 - \beta_T \beta\right) c\Delta t.$$

If $\beta_T \beta > 1$ as it is for our experimenters, then $c\Delta t'$ has the opposite sign of $c\Delta t$. The reverse is also true for the reply signal. We can conclude from this that the spacetime diagram is still a complete geometrical representation of the Lorentz transformation.

It gets even weirder when we consider the coordinate velocity transformation. In Joe's IRF, the tachyon travels away from him at $u = 5c$ toward Ashley. However, Ashley will measure the velocity of the tachyon to be

$$u' = \frac{u - v}{1 - uv/c^2} = \frac{\beta - \beta_T}{1 - \beta\beta_T} c = \frac{5 - 0.8}{1 - (5)\,(0.8)} c = -1.4c,$$

meaning the tachyon is traveling in the opposite direction! Because the tachyon is moving away from Joe faster than Ashley, we'd expect the tachyon to arrive and it does... but only in Joe's IRF. In Ashley's IRF, it never arrives because it's traveling away from Joe the other way! If it never arrives, then she can't send a reply and causality isn't violated. In other words, we can't draw the second blue arrow in Figure 7.21 because it never happens.

Matters get worse if we include a third IRF. Let's say another experimenter, Tiffany, is moving away from Joe with a relative velocity of $v = 0.2c$ (much more slowly than Ashley). The velocity she measures for the tachyon will be

$$u' = \frac{\beta - \beta_T}{1 - \beta\beta_T} c = \frac{5 - 0.2}{1 - (5)\,(0.2)} c,$$

which is undefined. The value of $v = 0.2c$ represents an infinite discontinuity of the coordinate velocity transformation. I say "undefined" rather than "infinite" because as $v \to 0.2c$ from lower speeds $u' \to +\infty$, but as $v \to 0.2c$ from higher speeds $u' \to -\infty$. That's two different extremes showing another mathematical problem. In general, this coordinate velocity boundary is $\beta_T = 1/\beta$ for any β_T and β. If $\beta_T < 1/\beta$, then the tachyon will be traveling the same direction in both frames (the two related by β_T). However, we don't have to worry about this creating a causality violation since

$$\beta_T < 1/\beta \quad \Rightarrow \beta_T \beta < 1$$

is the same condition which makes Δt and $\Delta t'$ both positive. Again, causality is maintained.

Mind you, this is all contingent on Joe being able to send information via tachyons. Recall that tachyons have imaginary mass and would have to interact with real mass to be sent by Joe. We're not even sure, given what we've learned so far in this book, how to physically interpret imaginary matter. It could very well not be capable of interacting with real matter in the first place. Remember, this is all speculative at this stage. We can only hope that some newer more advanced theory will explain these strange particles away.

7.7 Weirder Stuff: Paradoxes

The special theory of relativity is already weird. You might even think it can't possibly get any weirder than it already has. Unfortunately, it can get much weirder if you think about the possibilities or implications more deeply. One carefully constructed thought experiment could bring the entire theory crumbling down... or could it?

We call these **paradoxes** and they always turn out to be an indication of one of two things:

1. A false assumption given the nature of the model being used, or

2. That we've stepped beyond the scope of the model.

The former is usually due to some preconceived notion of how the universe functions based on our personal experience. All we have to do is let go of it and the problem disappears. The latter, on the other hand, is a bit more difficult to see. Sometimes it results from simplifying or idealizing the problem too much, which can be easily rectified. Other times, it can result from a lack of understanding with respect to the given conditions, which is much more difficult to resolve. Causality paradoxes, such as the one that resulted from the use of tachyons in Section 7.6, are a prime example of this. Carefully constructed problems require carefully constructed solutions.

In this section, we'll address a few well-known paradoxes and present their solutions.

Example 7.7.1

Two spaceships of equal proper length are traveling in opposite directions along the x-axis each with a constant relative speed of v. The ship traveling to the right is piloted by Joe and the other by Ashley. At the moment Ashley's ship's bow (or front end) lines up with Joe's ship's aft (or rear end), Ashley fires a laser from her ship's aft in an attempt to hit Joe's ship's bow. You may assume the ships are close enough together along the y-axis to neglect the travel time of the laser beam.

This presents a paradox if we think about the scenario in the context of special relativity. In Ashley's IRF, Joe's ship experiences a length contraction. That means she sees her laser miss Joe's ship because it's too short. In Joe's IRF, Ashley's ship experiences the length contraction. That means her laser will hit his ship somewhere toward the middle. Both events cannot occur, so which is it? A hit or a miss?

- We have seen that measurements taken in different IRFs are relative to the specific IRF. However, even though the measurements can be different, the events are not. If the laser fire misses in one IRF, then it must miss in all IRFs. The only difference is how and when it will miss. Similar reasoning applies if the laser fire hits.

- The paradox in this case is simply the result of a preconceived notion of time. It comes from the use of the phrase "at the moment" referring to Ashley firing the laser. Under classical physics, time is absolute and we need not worry about the subtleties. In the context of special relativity, however, this moment for Ashley may not be the same moment for Joe. We need to discuss this problem in terms of events and worldines.

- Ashley's detection of the aft of Joe's ship is an event in spacetime and the laser fire at the bow of Joe's ship is a completely separate event. This is shown in Figure 7.22 as events 1 and 2, respectively. You can see these events are simultaneous in Ashley's frame (the unprimed frame). Furthermore, from her perspective, the shot misses because Joe's ship is too short as expected from the example description.

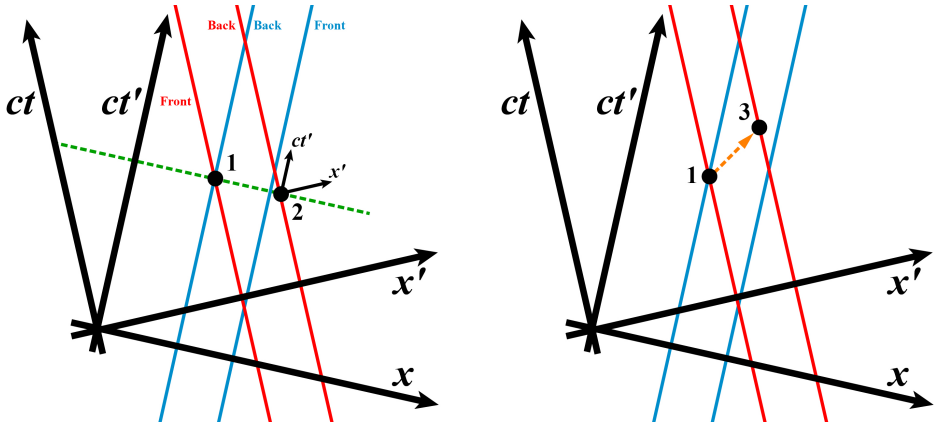


Figure 7.22: In this spacetime diagram, the two blue world lines correspond to the front and back of a spaceship moving to the right (and the red worldines, to the left). Event 1 is the detection of the back of the blue ship by the front of the red ship. Event 2 is the red ship firing a laser beam, which is simultaneous to event 1 in the unprimed frame. Event 3 is the firing of the same laser, but accounting for the time required for the signal to travel from the front of the red ship to the back telling the laser to fire.

- In Joe's frame, event 1 happens after event 2! That means, from his perspective, Ashley fired the shot before the ends of the ships line up (i.e. too early). For him, the shot also misses because the bow of his ship still hasn't lined up with the aft of hers. Events 1 and 2 are not the same moment for Joe. Both perspectives are shown in Figure 7.23.

- In fact, we can use a few numbers to see how far apart in time Joe measures these moments to be. Let's assume in Ashley's frame that events 1 and 2 occur at $(0,0,0,0)$ and $(0,50\,\text{m},0,0)$ meaning we've assumed the ship's proper length to be 50 meters. We'll also assume $v = 0.5c$ just for comparison. Using a Lorentz transformation (Eq. 7.3.1) on the spacetime coordinates, we get $(0,0,0,0) = (0,0,0,0)$ for event 1 since it's the zero vector and

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1.155 & -0.577 & 0 & 0 \\ -0.577 & 1.155 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 50\,\text{m} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -28.87\,\text{m} \\ 57.735\,\text{m} \\ 0 \\ 0 \end{bmatrix}$$

for event 2. The negative time component implies event 2 occurs $t = 28.87\,\text{m}/c = 96.3\,\text{ns}$ before event 1. This isn't much, but it's enough for

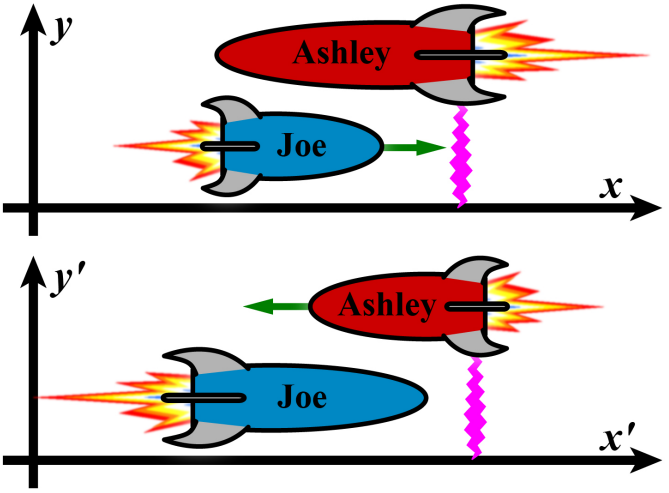


Figure 7.23: In the unprimed frame (Ashley's IRF), the laser fire (designated by the purple beam) misses because Joe's ship is too short due to length contraction. In the primed frame (Joe's IRF), the laser fire misses because the shot was fired too early.

the laser to miss Joe's ship. We can also see the laser misses Joe's ship by $57.735\,\text{m} - 50\,\text{m} = \boxed{7.735\,\text{m}}$. In Ashley's frame, the shot misses by

$$50\,\text{m} - \frac{50\,\text{m}}{\gamma} = 50\,\text{m} - \frac{50\,\text{m}}{1.155} = \boxed{6.7\,\text{m}},$$

but it still misses.

- We've solved the paradox by letting go of a preconceived notion of time. Unfortunately, it isn't a physically accurate solution because we didn't consider how the bow of Ashley's ship communicates with the aft of her ship. Assuming this communication is instantaneous is a physical impossibility because the fastest way to send information (under special relativity) is at c by, perhaps, a radio wave. We've taken this into account in Figure 7.22 by showing the signal as a orange dashed arrow. By the time the signal to fire reaches the laser weapon at the aft of Ashley's ship, both ships have moved enough so that the laser hits Joe's ship (in both IRFs).

Example 7.7.2

A common example for introductory students is the "pole in the barn" problem: A farmer holding a 6 meter long pole (perfectly horizontally) is running toward a small barn. If the barn is 5 meters from front to back and both the front and back doors are open, then how fast does the farmer have to run to fit the pole in the barn?

The idea is that the faster the farmer runs, the more contracted the length of the pole gets. If he runs fast enough, the pole should contract to the length of the barn. It's a relatively short calculation using length contraction (Eq. 7.2.12):

$$\gamma = \frac{L_{P,p}}{L_P} = \frac{L'_P}{L_P} = \frac{6\,\text{m}}{5\,\text{m}} = 1.2,$$

noting we only get to use this because one of the frames is the rest frame of the pole. We recall proper length, L_p, is defined as the maximum possible length measurement between two events. The two events in question are

1. The front of the pole lining up with the back of the barn and

2. The back of the pole lining up with the front of the barn.

According to Eq. 7.2.9, a value of $\gamma = 1.2$ corresponds to a velocity of $\beta = v/c = 0.5528$, a little over half the speed of light. That's totally unrealistic for the farmer, but not impossible in general.

- So where's the problem? The quantity L_P is the length of the pole as measured by the farmer's son who is stationary relative to the barn. The farmer running with the pole is still going to measure a length of 6 meters as shown in Figure 7.24. According to that same farmer, it is actually the barn that is moving at $\beta = 0.5528$, so the barn experiences the length contraction (Eq. 7.2.12):

  $$L'_B = \frac{L_{B,p}}{\gamma} = \frac{L_B}{\gamma} = \frac{5\,\text{m}}{1.2} = 4.167\,\text{m},$$

  noting proper length for the barn is measured in the unprimed frame (the barn's IRF). Only the farmer's son sees the pole fit in the barn. The farmer sees a minimum 6 m - 4.167 m = 1.833 m of the pole sticking out of the barn.

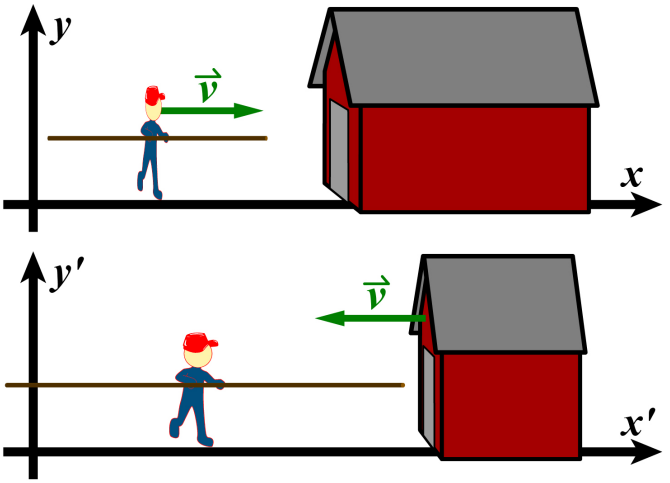


Figure 7.24: In the unprimed frame (the barn's IRF), the pole fits perfectly into the barn because the pole has length contracted. In the primed frame (the pole's IRF), the pole doesn't fit into the barn because the barn has length contracted.

- Still don't see a problem with this yet? That's ok because there really isn't a problem yet. Different observers measure different things all the time. In fact, we can use a Lorentz transformation (Eq. 7.3.1) assuming, in the farmer's son's frame, the spacetime coordinates are $(0, 5\,\text{m}, 0, 0)$ and $(0, 0, 0, 0)$ for events 1 and 2 respectively (i.e. the events are 5 meters apart and simultaneous). In the farmer's frame, event 2 becomes $(0, 0, 0, 0) = (0, 0, 0, 0)$ since it's the zero vector and event 1 becomes

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1.2 & -0.6633 & 0 & 0 \\ -0.6633 & 1.2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 5\,\text{m} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -3.317\,\text{m} \\ 6\,\text{m} \\ 0 \\ 0 \end{bmatrix},$$

where the negative time component implies event 1 occurs

$$t = \frac{3.317\,\text{m}}{c} = 11.06\,\text{ns}$$

before event 2 as shown in Figure 7.25. This makes perfect sense. If the pole doesn't fit in the barn, then the front will exit the barn before the back enters.

- Everything is just fine until the farmer's son decides to be a smart alec. What if he leaves the back door closed and, at the moment he sees the

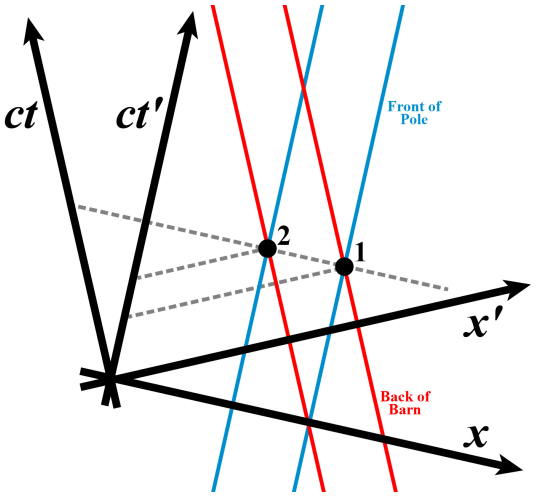


Figure 7.25: In this spacetime diagram, the two blue world lines correspond to the front and back of the pole (and, likewise, the red world lines to the barn). You can see those events are simultaneous in the unprimed frame (the barn's IRF). However, event 1 occurs before event 2 in the primed frame (the pole's IRF) as shown by the gray dashed lines.

back of the pole line up with the front of the barn (i.e. event 2), he closes the front door. According to the farmer, the pole doesn't fit, so is the pole in the barn or not?!

- We saw in Example 7.7.1 that the same set of events must occur in all frames of reference. Different frames just disagree on how, and sometimes in what order, those events unfold. If the pole is enclosed in the barn in the son's frame, then it must also be enclosed in the farmer's frame.
- In the farmer's frame, event 1 occurs 11.06 ns before event 2, so the doors don't close simultaneously for him, but that isn't quite enough to reconcile this paradox. We need to let go of one more thing: the rigidity of the pole.
- Since the back door is closed, it collides with the front of the pole. Assuming the door and the pole can survive the impact (which they probably can't) and the barn keeps moving at $\beta = 0.5528$ (which it probably isn't due to conservation of momentum), the barn door must start to move the front of the pole. However, the back of the pole doesn't notice and stays still because the speed

of light is the maximum speed at which information can travel. The pole experiences some extreme tensile stress. In the 11.06 ns it takes for the other barn door to close, the pole will have compressed by

$$vt = (0.5528c)\,(11.06\text{ ns}) = 1.833\text{ m}.$$

This is enough to fit in the barn: $6\text{ m} - 4.167\text{ m} = 1.833$ m.

– In short, the son sees the pole contract due to its motion and the farmer sees the pole contract due to tensile stress. Either way, the pole is enclosed in the barn ...at least for a moment or two until the pole is officially in the barn's frame. At that point, it's likely the pole and the barn doors will explode from the stress if they haven't already.

- A friend of mine once suggested another paradox that took me a while to resolve. He proposed a thought experiment avoiding the use of the barn doors. Suppose, attached to the pole, there is a battery and an LED connected in series with an open circuit on each end of the pole. A metal post is placed across the barn with connections hanging down to complete the circuit (See Figure 7.26). If the pole fits perfectly between the barn doors, then the LED will light. If not, the LED will not light.

  – You can't create a photon in one frame and not another. It must be created in all frames or none, with no exceptions. The problem in this case is we've stepped beyond the scope of the basic circuit model. The battery generates an electric field to move charges in a complete circuit. E-fields propagate at the speed of light, which *appears* instantaneous most of the time.

  – Unfortunately, since the pole is traveling at $\beta = 0.5528$, this propagation speed is no longer negligible. For the LED to light in the unprimed frame (the barn's IRF), the circuit must be complete for *at least*

$$t = \frac{L_B + L_P}{c} = \frac{5\text{ m} + 5\text{ m}}{3 \times 10^8\text{ m/s}} = 33.33\text{ ns}$$

to allow the E-field to propagate the round trip of the circuit. This is ignoring any response time the LED itself might need.

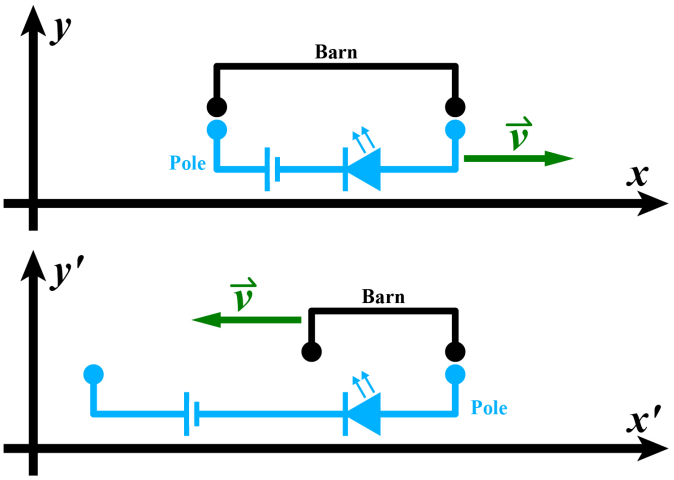


Figure 7.26: This is a representation of how event 1 appears to both observers in the pole-barn circuit paradox. In the unprimed frame (the barn's IRF), the contacts match up and the circuit is complete and the LED should light. In the primed frame (the pole's IRF), the circuit is not complete and the LED should not light.

– This may still seem like a very small amount of time, so let's consider it in context. In 33.33 ns traveling at $\beta = 0.5528$, the pole (or the barn) will have moved a distance of

$$\Delta x = vt = \beta ct = \beta \left(L_B + L_P \right) = 5.528 \text{ m},$$

so the circuit contacts must be *at least* this long. However, this is longer than the barn in either frame. The only way to make this distance negligible is to make β very small, which ultimately makes length contraction negligible and this entire conversation a moot point.

Example 7.7.3

Probably the most famous of all the paradoxes in special relativity is the "twin's paradox." The paradox itself stems from common problem given to introductory students. Here's the basic idea: You have a set of identical twins. One of them is an adventurous astronaut and the other a homebody.

On their 25th birthday, the astronaut hops in a spaceship and travels off to a star 8 ly away (let's say Wolf 359) at half the speed of light ($v = 0.5c$). Upon arriving at the star, the astronaut discovers nothing special and immediately heads home at the same speed.

The homebody twin observes her sister take 16 years to get to the star and another 16 years to get home. This makes sense since

$$v = \frac{\Delta x}{\Delta t} \quad \Rightarrow \Delta t = \frac{\Delta x}{v} = \frac{8\,c\,\text{yrs}}{0.5c} = 16\,\text{yrs}$$

for a one-way trip or 32 years for the roundtrip. That makes her now exactly 57 years old. However, due to time dilation (Eq. 7.2.11), the gamma factor (Eq. 7.2.9) is

$$\gamma = \frac{1}{\sqrt{1-\beta^2}} = \frac{1}{\sqrt{1-0.5^2}} = 1.155,$$

so the astronaut twin only experiences

$$\Delta t_p = \frac{\Delta t}{\gamma} = \frac{16\,\text{yrs}}{1.155} = 13.86\,\text{yrs}$$

for a one-way trip or 27.71 years for the roundtrip. This makes her only between 52 and 53 years old, 4–5 years younger than the homebody twin. All of this is perfectly legal in the context of special relativity as long as the two twins agree how old they each are.

The paradox here arises when we try to examine things from the astronaut's point of view. No frame of reference gets any preference over another, so the astronaut would consider herself stationary and the Earth moving at 0.5c. According to her, Earth has the shorter time. If the astronaut experiences a total of 27.71 years, then the homebody should experience

$$\Delta t_p = \frac{\Delta t}{\gamma} = \frac{27.71\,\text{yrs}}{1.155} = 24\,\text{yrs}$$

as opposed to 32 years. It would seem the twins do not agree on how much time has passed on Earth, so who is correct?

When considering the total time passed during the roundtrip, it turns out the Earth is correct about the Earth's time as you might expect. However, the reasoning behind why is far from straight forward. I've found a wide variety of explanations ranging from incomplete to unnecessarily complicated to just plain wrong. Here are some common examples:

1. "*The spaceship experiences acceleration, so it's beyond the capabilities of special relativity. You need general relativity to resolve the issue.*" Special relativity is perfectly capable of dealing with accelerating objects (see Example 7.3.2). It just can't deal with accelerated reference frames (ARFs) meaning we can't discuss anything the spaceship measures while accelerating without invoking general relativity (Chapter 8). Furthermore, what happens during the accelerating portions of the trip has no bearing on what happens during the uniform motion portions of the trip.

2. "*The reference frames are not symmetric because the spaceship experiences acceleration meaning it isn't an inertial reference frame. Since Earth is the only IRF, it gets preference.*" First off, any explanation like this is a cop-out because it dodges any discuss of real physics. Secondly, we can very easily stop and start the clocks to avoid including the acceleration in the problem entirely. Doing so does not resolve the paradox.

3. "*The twins cannot observe each other's clock without seeing light from each other, which takes time to travel between them.*" This statement is true and it might affect how we'd actually see the time pass between the beginning and the end. However, it is by no means a resolution to the twin's paradox. All observers agree on the speed of light, so we all know how long it takes and it can be factored out of our calculations. Some references on special relativity have even resorted to invoking Doppler effect, which even further complicates the situation.

4. "*When the spaceship turns around, it switches IRFs, which changes the lines of simultaneity for the spaceship but not the Earth.*" This one has some promise, but is severely incomplete. My guess is someone figured this out 100 years ago, but it's been copy/pasted so many times that we've forgotten what the point actually was. No one really understands it anymore (or at least the ones that do aren't talking about it).

To get at the real complete solution without getting lost, we're going to keep things as simple as possible by removing all unnecessary factors. First, we'll assume that both observers can account for light travel time and leave it out of the discussion. Second, we're going to remove all accelerations from the problem by only running the two clocks during constant velocity portions

of the trip. This will involve starting and stopping the clocks a couple times. Then finally, we're going to consider the two halves of the trip completely independently.

- We'll assume the spaceship has been given time to accelerate to its cruising speed of $0.5c$ before the clocks pass each other and are started. Both clocks clearly start together since this is represented by the same event (i.e. they happen at the same place and same time).

- The clocks are not stopped until the spaceship reaches its destination of Wolf 359 (8 ly away as measured from Earth). The spaceship maintains its cruising speed until it stops its clock so as to avoid including accelerations. Also, since we're not including any signals transmitted between them, both observes agreed before departure to stop each of their clocks at the appropriate time.

- According to Earth, it took the spaceship 16 years to arrive at the destination just as we calculated before, so that's when Earth stops its clock. When the spaceship stops its own clock, it shows 13.86 years also just as calculated before.

- Now we bring the spaceship to rest relative to Wolf 359 for a while and have the astronaut talk to her homebody sister to compare notes. They begin to argue over how much time they think passed on Earth during the trip because, at least while the clocks were running, they each think they were stationary and the other was moving. This discrepancy is easily resolved with spacetime diagram (our go-to solution throughout this chapter).

- In order to keep things as clear as possible, Figure 7.27 is done to scale. You can see from the lines of simultaneity (i.e. all events occurring at the same time) that their disagreement stems from when they each think the Earth *should* have stopped its clock, not when Earth *actually did* stop its clock. The astronaut thinks Earth should have stopped its clock 4 years early (as measured in the Earth's IRF) bringing the 16 years down to 12 years (half of the 24 years calculated earlier). There is no paradox because the clocks only stop at the same time in Earth's IRF.

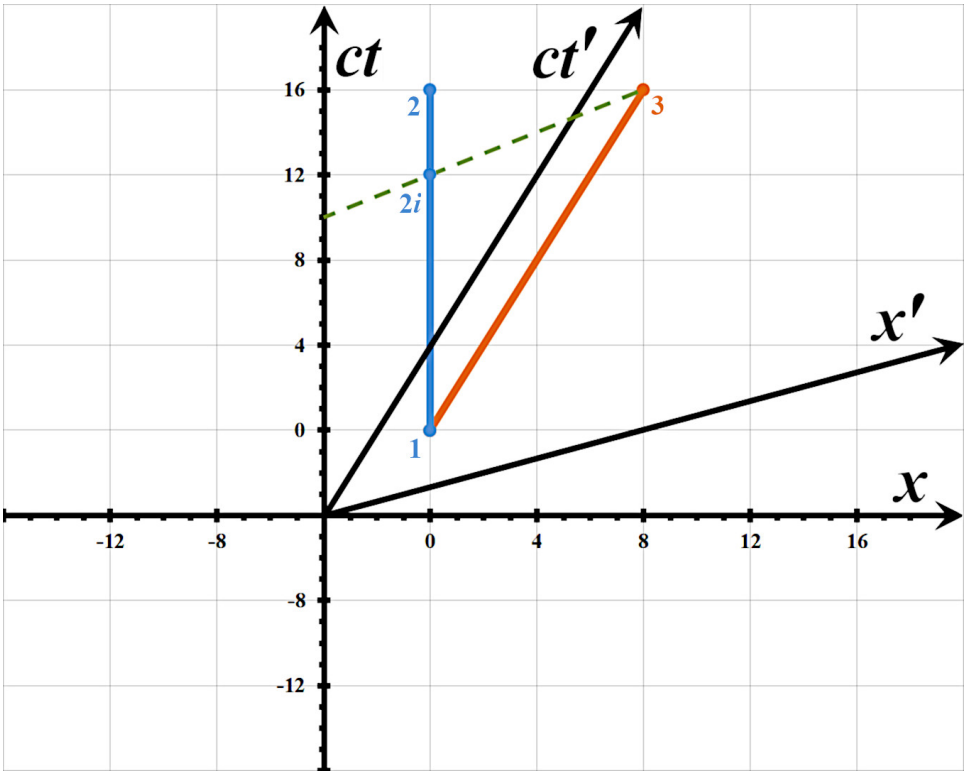


Figure 7.27: Two clocks start at event 1. An astronaut travels to the star Wolf 359 between events 1 and 3. Her twin sister stays on Earth traveling between events 1 and 2. Events 2 and 3 represent when each twin stops their clock, which only occurs at the same time in the unprimed frame (Earth's IRF). The green dashed line connects all the events happening simultaneously in the primed frame (spaceship's IRF). It is clear the astronaut thinks her twin should have stopped her clock after 12 years (at event $2i$) rather than after 16 years.

- Using the Lorentz transformation (Eq. 7.3.1) method, event 2 becomes

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1.155 & -0.5774 & 0 & 0 \\ -0.5774 & 1.155 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 16\,c\,\text{yrs} \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 18.48\,c\,\text{yrs} \\ -9.238\,c\,\text{yrs} \\ 0 \\ 0 \end{bmatrix},$$

  and event 3 becomes

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1.155 & -0.5774 & 0 & 0 \\ -0.5774 & 1.155 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 16\,c\,\text{yrs} \\ 8\,c\,\text{yrs} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 13.86\,c\,\text{yrs} \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

  In the spaceship's IRF, spaceship measures the distance between them as 9.238 ly (considering itself to be at zero). It also measures a 4.62 year difference between events 2 and 3, particularly that event 3 occurs 4.62 years before event 2. In Earth's IRF, that would be measured as

$$\frac{4.62\,\text{yrs}}{\gamma} = \frac{4.62\,\text{yrs}}{1.155} = 4\,\text{yrs},$$

  which is exactly what we got with the spacetime diagram method with way less work.

- On the return trip, the reverse happens as shown in Figure 7.28. After communicating a little more, they agree to start the clocks again at a designated time. However, those events again only occur simultaneously in the Earth's IRF, not the spaceship's. Upon arrival at Earth, the astronaut yells at her homebody sister for starting her clock 4 years too early.

- Now let's consider it all together. Notice the spaceship switches from a single-primed frame to a double-primed frame between Figures 7.27 and 7.28 because it switched directions. If it's still unclear, Figure 7.29 shows the whole trip. That's assuming the astronaut stays at the star for 16 years, enough time for one message and a response.

- Essentially, $\Delta t_{Earth} = 32$ years while Earth's clock is running. The $\Delta t = 24$ years calculated earlier for Earth involves a different set of four events, two of which ($2i$ and $4i$) are completely in the imagination of the astronaut. The Earth measures its own time correctly because it's controlling its own clock.

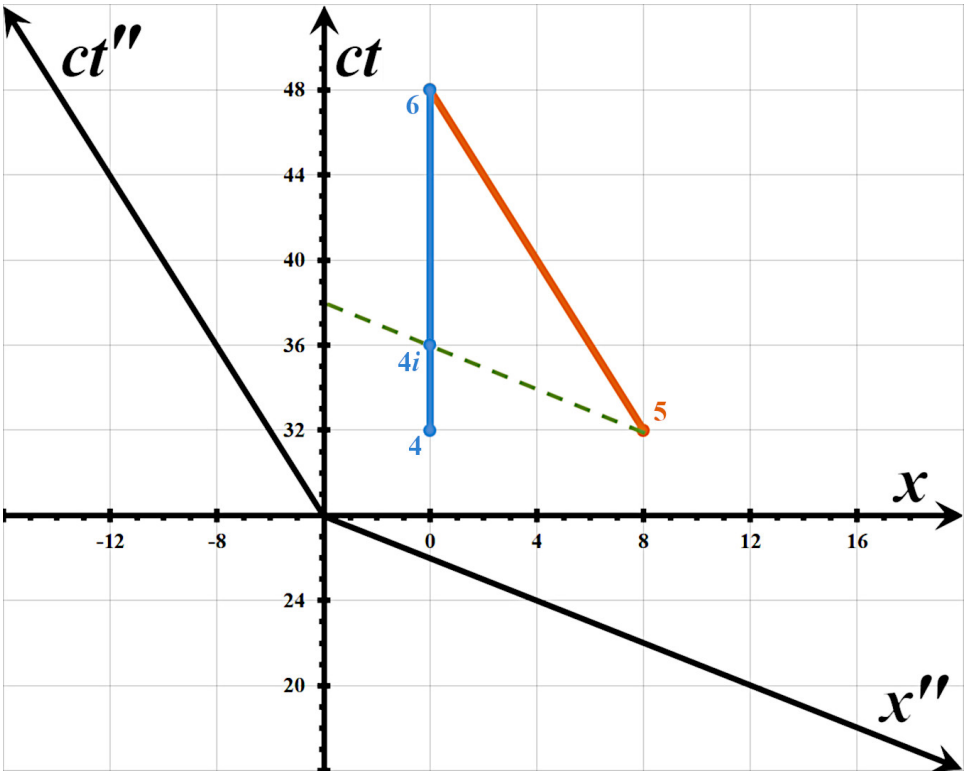


Figure 7.28: This is the trip home occurring after Figure 7.27. An astronaut travels home between events 5 and 6 while her twin sister on Earth travels between events 4 and 6. Events 4 and 5 represent when each twin restarts their clock, which only occurs at the same time in the unprimed frame (Earth's IRF). The green dashed line connects all the events happening simultaneously in the double-primed frame (spaceship's IRF). It is clear the astronaut thinks her twin should have started her clock 4 years later (at event $4i$).

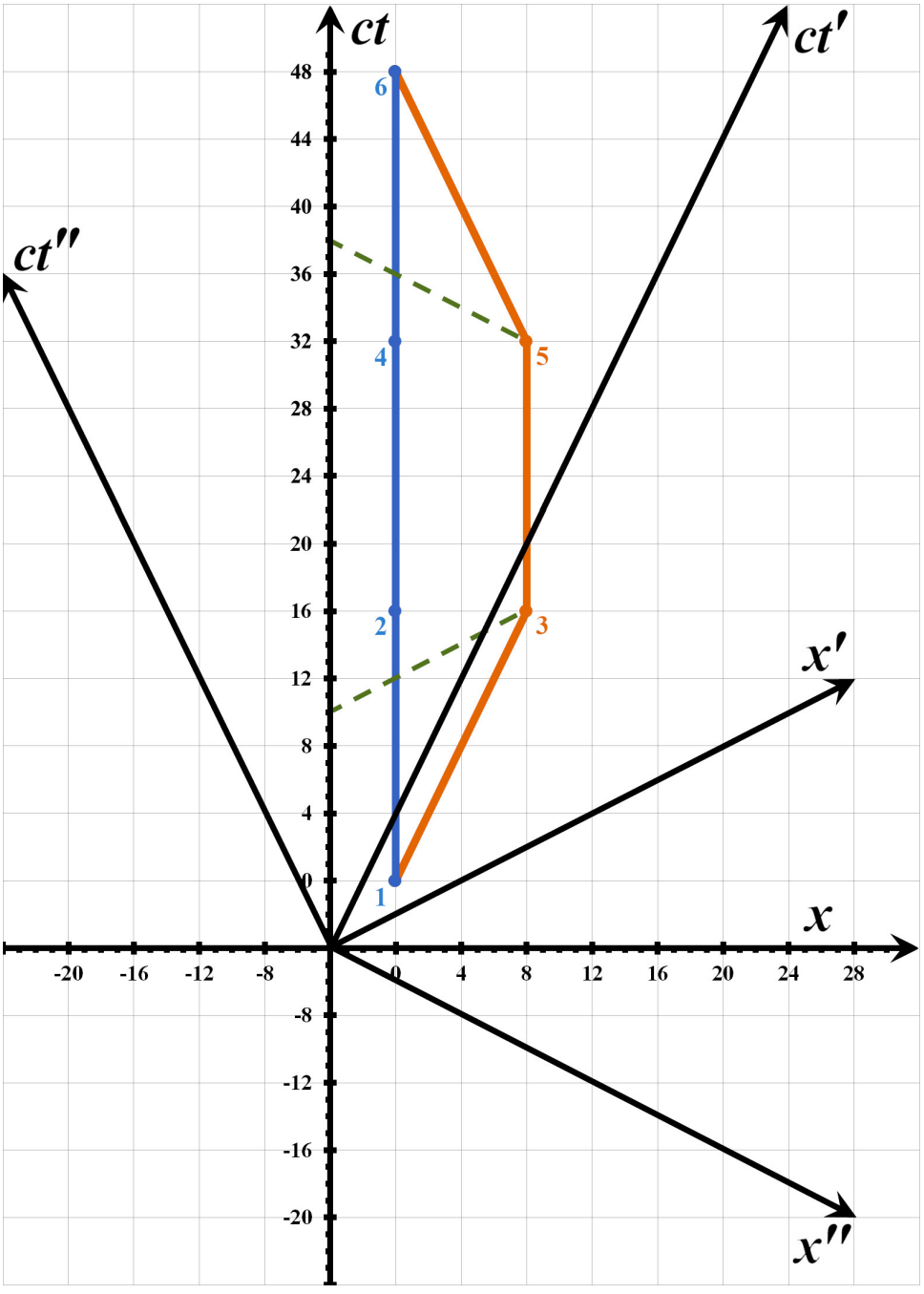


Figure 7.29: This is the entire trip from Figures 7.27 and 7.28 involving the two twins. It includes all three reference frames and all six real events.

The weirdest consequence shown in the Figures 7.27, 7.28, and 7.29 is how much time passes for each observer during the accelerations. These four accelerations are sharp corners, which means the acceleration occurs during a very short time period for the astronaut. The one-way trip is measured in *years* for both observers, so let's assume each acceleration only took two *days*. Yes, I'm aware that corresponds to a very violent proper acceleration (Eq. 7.4.17) of $88.5g$ (i.e. 88.5 times the gravity of Earth), which is far too high for any human to survive for two days straight. Unfortunately, a comfortable $1g$ would require 177 days (or about six months), which is far too long to ignore. Just go with it.

It doesn't get weird until we look at how the Earth sees the astronaut slow down at Wolf 359. According to Figure 7.27, the astronaut switches IRFs at event 3. By what we just assumed, event 3 is a two-day deceleration for the astronaut. The *beginning* of event 3 is simultaneous with event $2i$, but the *end* of event 3 is simultaneous with event 2 (since it's now in the rest frame of Earth). The time between event $2i$ and event 2 is four *years*! Truly understanding what happens during those accelerations would require general relativity (Chapter 8), but I have yet to see anyone use it to tackle this particular version of the paradox.

Chapter 8

General Relativity

8.1 Origins

Shortly after publishing his five papers in 1905, Albert Einstein began thinking a bit more about his theory of relativity. He had successfully ended the argument between classical mechanics and electrodynamics, which was certainly no small feat. However, the solution had one small limitation: it couldn't accurately predict measurements taken inside an accelerated reference frame (ARFs). This seems like a small issue, but always taking measurements in inertial reference frames can be occasionally inconvenient since the surface of the Earth is only *approximately* inertial (e.g. it rotates slowly). It also indicates a gap in our understanding and science has a drive to fill such gaps. Einstein knew he needed a more general theory of relativity (hence "general relativity"). This would involve at least one more postulate to address this issue, so he began performing more thought experiments.

Equivalence Principle

Explaining phenomena in an ARF can be tricky because of **fictitious forces** (i.e. forces that do not exist in all frames of reference.) The most popular examples of these are the Coriolis and centrifugal forces which exist in a rotating reference frame, but disappear in an inertial frame. The rotation itself is enough to explain the motion in the inertial frame. In 1907, Einstein's thoughts were on a much simpler type of ARF: a rocket accelerating in a straight line. He realized if a rocket accelerated at 9.8 $\mathrm{m/s^2}$ and its

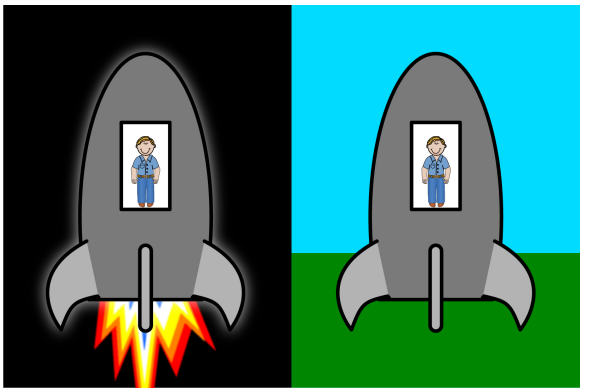


Figure 8.1: On the left, a rocket is accelerating through space at 9.8 m/s^2. On the right, an identical rocket is a rest on the surface of the Earth. These two situations are indistinguishable to the observers inside the rockets.

passengers were enclosed in a sound/shake-proof room with no windows, then the passengers would not be able to distinguish this motion from the gravitational field of the Earth (9.8 N/kg).

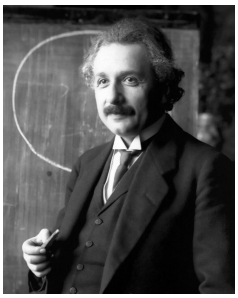
Einstein took this a step further, however. He postulated that these two phenomena were not just indistinguishable, but were in fact *equivalent*. The **equivalence principle**, as it has come to be called, is stated simply as

- When observing a behavior, whether it is caused by acceleration or by gravity is only a matter of reference frame. They are equivalent explanations.

What he meant was the fictitious force resulting from the acceleration is not fictitious at all. It is *literally* gravity! It would appear you can't explain acceleration without also explaining gravity in the same context. The ultimate implications of this were, at the time, beyond what anyone could foresee, but it got the wheels turning for Einstein and a few others.

Spacetime Revisited

As mentioned in Section 7.2, Hermann Minkowski generalized Einstein's work in 1908 by describing spacetime itself with tensor analysis. This got Einstein thinking about his equivalence principle a bit more. "What if spacetime is something tangible? What if it can be changed?" he asked himself. Not being


Albert Einstein


Marcel Grossmann


Tullio Levi-Civita


David Hilbert

Figure 8.2: These people were important in the development of general relativity.

particularly skilled in advanced mathematics (e.g. tensor analysis), he struggled for a few years. By 1912, he gave up and consulted a couple mathematicians, Marcel Grossmann and Tullio Levi-Civita, who recommended combining differential geometry and tensor analysis as the best possible method for finding a solution.

Unbeknownst to Einstein, a mathematician named David Hilbert (a very close friend of Minkowski's) was also working on the same problem using the same methods. It wasn't until the summer of 1915, when Hilbert invited Einstein to the Göttingen Mathematics Institute to give several lectures on his recent work, that Einstein learned about Hilbert's work. You might think this would raise tensions between the two men, but there is no historical indication of this. Einstein and Hilbert began consulting each other between July and November of that year, both publishing small papers along the way. This ultimately resulted in full papers being published almost simultaneously by each of them describing the nature of spacetime and gravity.

Spacetime Curves?!

We mentioned the use of something called differential geometry, which is very important in the development of general relativity. It's a mathematical tool describing the behavior of not only curves, but surfaces and volumes as well. The way it's formulated allows it to apply to any number of dimensions, including but not limited to the four-dimensional spacetime in which we live. It's common to think of spacetime as a "fabric" of sorts that can be stretched, compressed, bent, twisted, etc. The more that fabric is deformed, the more energy it contains and, therefore, the more it can influence anything

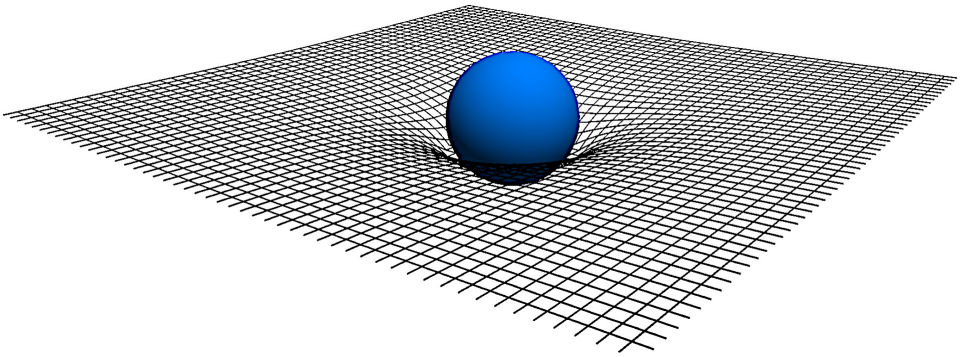


Figure 8.3: A common visual curved spacetime is the rubber sheet analogy, featured here. If we rolled a marble across this mesh sheet, then it would be drawn to the ball in the center. Unfortunately, spacetime doesn't *actually* look like this, so it's only good for demonstrating the concept of curvature. We'll develop a much more accurate diagram later in Section 8.6.

in contact with it.

For a linear curve, the **curvature** involves only one number at every point along the curve: the second derivative of the curve at that point. We've actually done this before when describing the behavior of waves (Eq. 5.5.3). It's not difficult to generalize this visual to a little further to a surface (see Figure 8.3). Unfortunately, spacetime fabric is four-dimensional, not one-dimensional nor two-dimensional. Our description of its curvature will require something called a **Riemann curvature tensor**,

$$R^{\delta}_{\alpha\mu\nu} = \frac{\partial \Gamma^{\delta}_{\alpha\nu}}{\partial x^{\mu}} - \frac{\partial \Gamma^{\delta}_{\alpha\mu}}{\partial x^{\nu}} + \Gamma^{\delta}_{\lambda\mu}\Gamma^{\lambda}_{\alpha\nu} - \Gamma^{\delta}_{\lambda\nu}\Gamma^{\lambda}_{\alpha\mu}\,, \tag{8.1.1}$$

which is a rank-4 dimension-4 mixed tensor (see Section 6.2 for more details on rank and dimension). This tensor isn't perfectly symmetric, but its last two indices obey

$$R^{\delta}_{\alpha\mu\nu} = -R^{\delta}_{\alpha\nu\mu}, \tag{8.1.2}$$

which is called skew symmetry. If you make the Riemann curvature tensor completely covariant, then we get

$$R_{\lambda\alpha\mu\nu} = -R_{\alpha\lambda\mu\nu} = -R_{\lambda\alpha\nu\mu}, \tag{8.1.3}$$

where $R_{\lambda\alpha\mu\nu} = g_{\lambda\delta}R^{\delta}_{\alpha\mu\nu}$ (note: index order is important). Also, performing this index operation multiple times can switch the sign back to positive (e.g. $R_{\lambda\alpha\mu\nu} = R_{\alpha\lambda\nu\mu}$ or $R_{\lambda\alpha\mu\nu} = R_{\mu\nu\lambda\alpha}$).

Because of the many ways a four-dimensional "fabric" can be deformed, every point in spacetime is assigned $4^4 = 256$ numbers (4 indices, each with a possible 4 values) to represent the total curvature. Notice the Riemann curvature tensor involves the Christoffel symbols (Eq. 6.7.6), which described the parallel transport of tensors during covariant derivatives (Eq. 6.7.5). Since the Riemann tensor describes curvature, it's actually a second derivative (i.e. $\nabla_\alpha\nabla_\delta T^{\mu\nu}$ for an arbitrary tensor $T^{\mu\nu}$) and so involves the product of two Christoffel symbols rather than just one. We can also take covariant derivatives of the Riemann tensor and get some useful identities. One is called a Bianchi identity,

$$\nabla_\sigma R_{\lambda\alpha\mu\nu} + \nabla_\lambda R_{\alpha\sigma\mu\nu} + \nabla_\alpha R_{\sigma\lambda\mu\nu} = 0, \tag{8.1.4}$$

where we essentially have even permutations of the first three indices.

Fortunately, a complete description of gravity doesn't require 256 values at every point. We can reduce (or "contract") the Riemann curvature tensor to two indices by summing over the other two, $R^{\mu}_{\alpha\mu\nu} = g^{\mu\lambda}R_{\lambda\alpha\mu\nu}$. This results in the **Ricci curvature tensor**,

$$R_{\alpha\nu} = \frac{\partial\Gamma^{\mu}_{\alpha\nu}}{\partial x^\mu} - \frac{\partial\Gamma^{\mu}_{\alpha\mu}}{\partial x^\nu} + \Gamma^{\mu}_{\lambda\mu}\Gamma^{\lambda}_{\alpha\nu} - \Gamma^{\mu}_{\lambda\nu}\Gamma^{\lambda}_{\alpha\mu}, \tag{8.1.5}$$

containing $4^2 = 16$ numbers. Furthermore, the Ricci tensor is symmetric (i.e. $R_{\alpha\nu} = R_{\nu\alpha}$), so this turns out to really be only 10 independent numbers. Contracting again gives us the **Ricci curvature scalar**,

$$R = R^{\nu}_{\nu} = g^{\alpha\nu}R_{\alpha\nu}, \tag{8.1.6}$$

which may come in handy since energy is a scalar quantity. The Ricci scalar contains less information than the Ricci tensor, so we'll need both as we describe the behavior of spacetime.

8.2 Einstein's Equation

The way physics handles derivations can be sneaky, but it can also save us a bit of time. In fact, this derivation is more of an argument than a derivation.

If you're looking for a more mathematically rigorous derivation, see Section 8.3.

First, we know that whatever result we get must approach the classical description at the classical limit (i.e. when the gravity field is weak and particles move slowly). Gravity is classically described using potential and mass density through Poisson's equation (Eq. 5.6.5) most well-known for its electrodynamics applications. For gravity, this is

$$\nabla^2\phi = 4\pi G\rho, \tag{8.2.1}$$

where the information about the gravity field is on the left and the matter on the right ($G = 6.674 \times 10^{-11}$ Nm^2/kg^2 is the gravitational constant). Whatever general equation we derive must be consistent with this.

If we're going to generalize using tensors, then the choice that comes to mind for the matter is the **stress-energy tensor**, $T_{\alpha\nu}$. This was briefly described in matrix form in Section 6.3 as

$$T_{\alpha\nu} \longrightarrow \begin{bmatrix} T_{00} & T_{01} & T_{02} & T_{03} \\ T_{10} & T_{11} & T_{12} & T_{13} \\ T_{20} & T_{21} & T_{22} & T_{23} \\ T_{30} & T_{31} & T_{32} & T_{33} \end{bmatrix}.$$

with it's various components having meanings and units that are unimportant for the time being. We will address them later. We should note though that this tensor is symmetric (i.e. $T_{\alpha\nu} = T_{\nu\alpha}$) just like the Ricci curvature tensor, so it also only contains 10 independent numbers. It also contains everything we could possibly want to know about the matter in the region.

Given that the stress-energy tensor and the Ricci curvature tensor behave in similar ways, it seems like the logical first try at a general equation would be

$$R_{\alpha\nu} = \kappa\, T_{\alpha\nu}, \tag{8.2.2}$$

where κ is some unknown constant we will determine later. Unfortunately, this violates a tensor form of the principle of conservation of energy:

$$\nabla^\alpha T_{\alpha\nu} = 0\,, \tag{8.2.3}$$

where $T_{\alpha\nu}$ is the stress-energy tensor and $\nabla^\alpha = g^{\alpha\lambda}\nabla_\lambda$ (possible because $\nabla_\lambda g^{\alpha\delta} = 0$). See Section 8.4 for more details on the stress-energy tensor. By Eq. 8.2.2, this also says

$$\nabla^\alpha R_{\alpha\nu} = 0.$$

On the other hand, by reducing with the Bianchi identity (Eq. 8.1.4), we get

$$g^{\nu\sigma}g^{\mu\lambda}\nabla_\sigma R_{\lambda\alpha\mu\nu} + g^{\nu\sigma}g^{\mu\lambda}\nabla_\lambda R_{\alpha\sigma\mu\nu} + g^{\nu\sigma}g^{\mu\lambda}\nabla_\alpha R_{\sigma\lambda\mu\nu} = 0$$

$$\nabla^\nu R_{\alpha\nu} + \nabla^\mu R_{\alpha\mu} - \nabla_\alpha R = 0,$$

because $R_{\alpha\nu} = R^\mu_{\alpha\mu\nu} = g^{\mu\lambda}R_{\lambda\alpha\mu\nu}$ and index order matters because of skew symmetry (i.e. $R_{\sigma\lambda\mu\nu} = -R_{\sigma\lambda\nu\mu}$). Since the summation index can change symbols on a whim, the first two terms are the same and this reduces to

$$\nabla^\mu R_{\alpha\mu} = \frac{1}{2}\nabla_\alpha R, \tag{8.2.4}$$

which implies R is constant (since its derivative is zero). This is troublesome since it means the curvature of spacetime is constant and, by Eq. 8.2.2, that T (the matter-energy distribution) is also constant throughout the entire universe.

Given that our universe does *not* have uniform density, we'll need a better option. The easiest way to handle this is to just add a second unknown term to the left side of Eq. 8.2.2,

$$R_{\alpha\nu} + X_{\alpha\nu} = \kappa\, T_{\alpha\nu}, \tag{8.2.5}$$

where we just need to solve for $X_{\alpha\nu}$. By conservation of energy (Eq. 8.2.3), this is

$$\nabla^\alpha R_{\alpha\nu} + \nabla^\alpha X_{\alpha\nu} = 0.$$

By Eq. 8.2.4, we get

$$\frac{1}{2}\nabla_\alpha R + \nabla^\alpha X_{\alpha\nu} = 0$$

$$\nabla^\alpha X_{\alpha\nu} = -\frac{1}{2}\nabla_\nu R$$

$$\nabla^\alpha X_{\alpha\nu} = -\frac{1}{2}g_{\alpha\nu}\nabla^\alpha R.$$

Since the covariant derivative of the metric is always zero ($\nabla^\alpha g_{\alpha\nu} = g_{\alpha\lambda}\nabla_\lambda g_{\alpha\nu} = 0$), this becomes

$$\nabla^\alpha X_{\alpha\nu} = \nabla^\alpha \left(-\frac{1}{2} g_{\alpha\nu} R\right)$$

$$X_{\alpha\nu} = -\frac{1}{2} g_{\alpha\nu} R,$$

assuming we're not adding any constants into the mix. Historical note: In 1922, Einstein tried to add a constant term to keep the universe static in size. He called it the cosmological constant... and then later called it the "biggest blunder" of his career. We will not be including such a constant.

If we substitute this back into Eq. 8.2.5, we get

$$R_{\alpha\nu} - \frac{1}{2} g_{\alpha\nu} R = \kappa\, T_{\alpha\nu}.$$

If we want this to reduce to Eq. 8.2.1 in the **weak-field approximation**, then $\kappa = 8\pi G/c^4$ and the final result is called **Einstein's equation**,

$$R_{\alpha\nu} - \frac{1}{2} g_{\alpha\nu} R = \frac{8\pi G}{c^4}\, T_{\alpha\nu}\,. \tag{8.2.6}$$

Sometimes this is called "Einstein's field equations" because there are actually 10 equations, one for each possible independent component of the tensors. It should also be noted that Einstein's equation is defined at a single arbitrary position in spacetime (i.e. an event) just like divergence and curl (see Section 3.2).

8.3 Hilbert's Approach

Einstein and Hilbert, coming from very different backgrounds, has very different ways of looking at problems. The method of choice for a mathematician like Hilbert was to start with a fundamental definition and work out every little detail until a solution. It's best to start this derivation with a quantity we only briefly mentioned near the end of Section 7.4. This quantity is called an **action**, which is a scalar field (i.e. a collection of scalars at various

points in space) like electric potential. However, an action is a measure of the efficiency of a path in spacetime and is defined as

$$S(q) \equiv \int_{t_1}^{t_2} \mathcal{L}(q, \dot{q})\ dt, \tag{8.3.1}$$

which is a line (or path) integral of the Lagrangian, $\mathcal{L}$, between times t_1 and t_2. Recall from Section 4.2, the Lagrangian is defined as the kinetic energy minus the potential energy and has standard energy units. As a result, in SI units, the action is measured in joule seconds (J·s).

The **principle of stationary action** states that an object or a particle will take a path with no variation in its action. We use the word "stationary" to mean zero variation like what occurs at a maximum or minimum (or saddle point on curved surfaces). In mathematical terms, we say

$$\delta S = 0, \tag{8.3.2}$$

where the delta operates on the action, S, to give us the variation. This is sometimes viewed as an alternate form of Lagrange's equation (Eq. 4.2.14) since they both involve the Lagrangian and both give the path taken.

If we intend on using the principle of stationary action in general relativity, then we'll have a generalize the definition for an action first. Rather than being integrated over just time, it should be over all spacetime. Also, if we include spacial coordinates, then we'll need a Jacobian multiplier (see Example 6.6.1) for the spacetime volume element.

$$S \equiv \int \mathcal{L}_{\text{total}} \sqrt{|\det(g)|}\, d^4x, \tag{8.3.3}$$

where g is the metric tensor in matrix form. Keep in mind, from here on out, we're sticking with the traditional sign convention for components of the metric tensor: $(-,+,+,+)$ initially defined in Section 7.2.

When writing the total Lagrangian for the system, it isn't enough to know about the matter in the region. In Section 7.5, we examined the relativistic nature of the electromagnetic field, which contains energy. As a result, the electromagnetic Lagrangian is

$$\mathcal{L}_{\text{EM}} = \frac{1}{4\mu_0} \mathcal{F}_{\alpha\delta} \mathcal{F}^{\alpha\delta} \tag{8.3.4}$$

where $\mathcal{F}_{\alpha\delta}$ is the electromagnetic field tensor given by Eq. 7.5.13. In fact, the tensor product above is given by Eq. 7.5.15.

Now that spacetime itself is a tangible entity, it too can have energy. Therefore, the total Lagrangian is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{matter}} + \mathcal{L}_{\text{EM}} + \mathcal{L}_{\text{spacetime}}.$$

However, since we're only interested in how spacetime and matter interact, we'll ignore the electromagnetic field for now. That means

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{matter}} + \mathcal{L}_{\text{spacetime}},$$

The spacetime Lagrangian can be written as

$$\mathcal{L}_{\text{spacetime}} = \frac{R}{2\kappa} = \frac{c^4}{16\pi G} R, \tag{8.3.5}$$

where κ is just a constant (consistent with Section 8.2). Note that the spacetime Lagrangian is zero when the curvature is zero. This is physically important and totally consistent with our "fabric" analogy. If you'd like to add a cosmological constant like the one mentioned in Section 8.2, then you'd add it here by giving flat spacetime a non-zero energy.

We are now in a position to be applying the principle of stationary action (Eq. 8.3.2). The total action can be written from Eq. 8.3.3 as

$$S = \int \left(\mathcal{L}_{\text{matter}} + \mathcal{L}_{\text{spacetime}}\right) \sqrt{|\det(g)|}\, d^4x$$

$$S = \int \left(\mathcal{L} + \frac{R}{2\kappa}\right) \sqrt{|\det(g)|}\, d^4x,$$

where $\mathcal{L} \equiv \mathcal{L}_{\text{matter}}$. Taking the variation of this action and applying the principle of stationary action, we get

$$0 = \delta \int \left(\mathcal{L} + \frac{R}{2\kappa}\right) \sqrt{|\det(g)|}\, d^4x$$

$$0 = \int \delta \left[\left(\mathcal{L} + \frac{R}{2\kappa}\right) \sqrt{|\det(g)|}\right] d^4x.$$

The variation operator works just like a derivative, so by the chain rule (Eq. 3.1.2)

$$0 = \int \frac{\delta}{\delta g^{\alpha\nu}} \left[\left(\mathcal{L} + \frac{R}{2\kappa} \right) \sqrt{|\det(g)|} \right] \delta g^{\alpha\nu} d^4x$$

Since this statement should be true for any variation in the inverse metric, $g^{\alpha\nu}$, we get

$$0 = \frac{\delta}{\delta g^{\alpha\nu}} \left[\left(\mathcal{L} + \frac{R}{2\kappa} \right) \sqrt{|\det(g)|} \right]$$

Don't get cancel-happy! Remember, this isn't *actually* a derivative. It's a variation, so integrating wont undo the operation. We have to evaluate it as is.

The variation works similar enough to a derivative to use the product rule (Eq. 3.1.5), so the variation becomes

$$0 = \sqrt{|\det(g)|} \frac{\delta}{\delta g^{\alpha\nu}} \left(\mathcal{L} + \frac{R}{2\kappa} \right) + \left(\mathcal{L} + \frac{R}{2\kappa} \right) \frac{\delta}{\delta g^{\alpha\nu}} \sqrt{|\det(g)|}. \tag{8.3.6}$$

Let's take a closer look at the variation in the second term. We know $\sqrt{|\det(g)|}$ is the same as $\sqrt{-\det(g)}$ in spacetime (i.e. you either have one negative or three negatives by convention), so

$$\delta \sqrt{|\det(g)|} = \delta \sqrt{-\det(g)} = -\frac{\delta \left[\det(g)\right]}{2\sqrt{-\det(g)}}.$$

We also know derivatives of determinants are given by the Jacobi formula,

$$\delta \left[\det(g)\right] = \det(g)\, g^{\alpha\nu}\, \delta g_{\alpha\nu} = -\det(g)\, g_{\alpha\nu}\, \delta g^{\alpha\nu}, \tag{8.3.7}$$

where we've taken advantage of $0 = \delta(g^{\alpha\nu} g_{\alpha\nu}) = g^{\alpha\nu}\, \delta g_{\alpha\nu} + g_{\alpha\nu}\, \delta g^{\alpha\nu}$. This means

$$\delta \sqrt{|\det(g)|} = -\frac{-\det(g)\, g_{\alpha\nu}\, \delta g^{\alpha\nu}}{2\sqrt{-\det(g)}} = -\frac{1}{2}\sqrt{-\det(g)}\, g_{\alpha\nu}\, \delta g^{\alpha\nu}$$

$$\frac{\delta}{\delta g^{\alpha\nu}} \sqrt{|\det(g)|} = -\frac{1}{2}\sqrt{-\det(g)}\, g_{\alpha\nu} = -\frac{1}{2}\sqrt{|\det(g)|}\, g_{\alpha\nu}.$$

If we substitute this into Eq. 8.3.6, then we get

$$0 = \sqrt{|\det(g)|}\,\frac{\delta}{\delta g^{\alpha\nu}}\left(\mathcal{L} + \frac{R}{2\kappa}\right) + \left(\mathcal{L} + \frac{R}{2\kappa}\right)\left(-\frac{1}{2}\sqrt{|\det(g)|}\,g_{\alpha\nu}\right)$$

$$0 = \frac{\delta}{\delta g^{\alpha\nu}}\left(\mathcal{L} + \frac{R}{2\kappa}\right) + \left(\mathcal{L} + \frac{R}{2\kappa}\right)\left(-\frac{1}{2}g_{\alpha\nu}\right).$$

Not looking familiar yet? This becomes

$$0 = -\frac{1}{2}\left(-2\frac{\delta\mathcal{L}}{\delta g^{\alpha\nu}} + g_{\alpha\nu}\mathcal{L}\right) + \frac{1}{2\kappa}\left(\frac{\delta R}{\delta g^{\alpha\nu}} - \frac{1}{2}g_{\alpha\nu}R\right)$$

when we combine like terms and pull out common factors.

This looks a little closer to what we want, but it needs a little work. We can simplify a bit more by moving terms to the other side, arriving at

$$\frac{\delta R}{\delta g^{\alpha\nu}} - \frac{1}{2}g_{\alpha\nu}R = \kappa\left(-2\frac{\delta\mathcal{L}}{\delta g^{\alpha\nu}} + g_{\alpha\nu}\mathcal{L}\right).$$

The parenthetical statement depends on the Lagrangian for the matter, $\mathcal{L}$, so must be related in some way to the stress-energy tensor, $T_{\alpha\nu}$. Upon close inspection, we can see it's both symmetric and conserved, so it must be proportional to $T_{\alpha\nu}$ (i.e. varies only by a constant coefficient). This coefficient would only take care of the units, but recall from our original description of $T_{\alpha\nu}$ in Section 8.2 that we're addressing issues with units later. We'll, therefore, go out on a limb to say the parenthetical quantity is *equal* to $T_{\alpha\nu}$.

As a result of incorporating the stress-energy tensor, the full equation becomes

$$\frac{\delta R}{\delta g^{\alpha\nu}} - \frac{1}{2}g_{\alpha\nu}R = \kappa\,T_{\alpha\nu} \tag{8.3.8}$$

and we can see we're almost there! There is just one variation left to evaluate: δR. Based on the definition of the Ricci curvature scalar (Eq. 8.1.6) and the product rule (Eq. 3.1.5), this means

$$\delta R = \delta(g^{\alpha\nu}R_{\alpha\nu}) = \delta g^{\alpha\nu}R_{\alpha\nu} + g^{\alpha\nu}\delta R_{\alpha\nu}$$

or, better yet,

$$\frac{\delta R}{\delta g^{\alpha\nu}} = R_{\alpha\nu} + g^{\alpha\nu}\frac{\delta R_{\alpha\nu}}{\delta g^{\alpha\nu}}.$$

The second term vanishes leaving just $R_{\alpha\nu}$ and Eq. 8.3.8 becomes

$$R_{\alpha\nu} - \frac{1}{2} g_{\alpha\nu} R = \kappa\, T_{\alpha\nu}\,, \tag{8.3.9}$$

which is exactly the result we got for Einstein's equation in Section 8.2.

What was that? Why does the second term vanish?! That was pretty blatant hand-waving, wasn't it? Explaining it, though, is going to take a little bit of careful planning. Remember the Ricci tensor is just a contraction of the Riemann tensor, so we'll avoid getting lost in the summation indices by starting with Riemann. Using the definition (Eq. 8.1.1), we get

$$\delta R^{\rho}_{\alpha\mu\nu} = \delta\left(\frac{\partial \Gamma^{\rho}_{\alpha\nu}}{\partial x^{\mu}} - \frac{\partial \Gamma^{\rho}_{\alpha\mu}}{\partial x^{\nu}} + \Gamma^{\rho}_{\lambda\mu}\Gamma^{\lambda}_{\alpha\nu} - \Gamma^{\rho}_{\lambda\nu}\Gamma^{\lambda}_{\alpha\mu} \right)$$

$$\delta R^{\rho}_{\alpha\mu\nu} = \frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\nu}\right)}{\partial x^{\mu}} - \frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\mu}\right)}{\partial x^{\nu}} + \delta\left(\Gamma^{\rho}_{\lambda\mu}\Gamma^{\lambda}_{\alpha\nu}\right) - \delta\left(\Gamma^{\rho}_{\lambda\nu}\Gamma^{\lambda}_{\alpha\mu}\right).$$

Using the product rule (Eq. 3.1.5) on the last two terms gives

$$\delta R^{\rho}_{\alpha\mu\nu} = \frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\nu}\right)}{\partial x^{\mu}} - \frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\mu}\right)}{\partial x^{\nu}} + \delta\Gamma^{\rho}_{\lambda\mu}\Gamma^{\lambda}_{\alpha\nu} + \Gamma^{\rho}_{\lambda\mu}\delta\Gamma^{\lambda}_{\alpha\nu} - \delta\Gamma^{\rho}_{\lambda\nu}\Gamma^{\lambda}_{\alpha\mu} - \Gamma^{\rho}_{\lambda\nu}\delta\Gamma^{\lambda}_{\alpha\mu}.$$

Moving some terms around and making sure the variations are always last, this is

$$\delta R^{\rho}_{\alpha\mu\nu} = \frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\nu}\right)}{\partial x^{\mu}} + \Gamma^{\rho}_{\lambda\mu}\delta\Gamma^{\lambda}_{\alpha\nu} - \delta\Gamma^{\rho}_{\lambda\nu}\Gamma^{\lambda}_{\alpha\mu} - \frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\mu}\right)}{\partial x^{\nu}} - \Gamma^{\rho}_{\lambda\nu}\delta\Gamma^{\lambda}_{\alpha\mu} + \delta\Gamma^{\rho}_{\lambda\mu}\Gamma^{\lambda}_{\alpha\nu}$$

$$\delta R^{\rho}_{\alpha\mu\nu} = \frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\nu}\right)}{\partial x^{\mu}} + \Gamma^{\rho}_{\lambda\mu}\delta\Gamma^{\lambda}_{\alpha\nu} - \Gamma^{\lambda}_{\alpha\mu}\delta\Gamma^{\rho}_{\lambda\nu} - \frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\mu}\right)}{\partial x^{\nu}} - \Gamma^{\rho}_{\lambda\nu}\delta\Gamma^{\lambda}_{\alpha\mu} + \Gamma^{\lambda}_{\alpha\nu}\delta\Gamma^{\rho}_{\lambda\mu}$$

and grouping gives us

$$\begin{aligned} \delta R^{\rho}_{\alpha\mu\nu} \;=\; & \left(\frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\nu}\right)}{\partial x^{\mu}} + \Gamma^{\rho}_{\lambda\mu}\delta\Gamma^{\lambda}_{\alpha\nu} - \Gamma^{\lambda}_{\alpha\mu}\delta\Gamma^{\rho}_{\lambda\nu} \right) \\ & - \left(\frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\mu}\right)}{\partial x^{\nu}} + \Gamma^{\rho}_{\lambda\nu}\delta\Gamma^{\lambda}_{\alpha\mu} - \Gamma^{\lambda}_{\alpha\nu}\delta\Gamma^{\rho}_{\lambda\mu} \right). \end{aligned}$$

Lastly, we can do some *voodoo math* (with a little foresight; we can add zeros, multiply by ones, add and subtract constants, etc. to simplify a mathematical expression) by subtracting a new term from the first parenthetical expression and adding that same term to the second. This results in

$$\begin{aligned}\delta R^{\rho}_{\alpha\mu\nu} &= \left(\frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\nu}\right)}{\partial x^{\mu}} + \Gamma^{\rho}_{\lambda\mu}\delta\Gamma^{\lambda}_{\alpha\nu} - \Gamma^{\lambda}_{\alpha\mu}\delta\Gamma^{\rho}_{\lambda\nu} - \Gamma^{\lambda}_{\mu\nu}\delta\Gamma^{\rho}_{\lambda\alpha}\right) \\ &\quad - \left(\frac{\partial\left(\delta\Gamma^{\rho}_{\alpha\mu}\right)}{\partial x^{\nu}} + \Gamma^{\rho}_{\lambda\nu}\delta\Gamma^{\lambda}_{\alpha\mu} - \Gamma^{\lambda}_{\alpha\nu}\delta\Gamma^{\rho}_{\lambda\mu} - \Gamma^{\lambda}_{\mu\nu}\delta\Gamma^{\rho}_{\lambda\alpha}\right).\end{aligned}$$

were the new term is $\Gamma^{\lambda}_{\mu\nu}\delta\Gamma^{\rho}_{\lambda\alpha}$.

A clever eye will recognize each of these parenthetical statements as covariant derivatives. Unlike the definition given in Eq. 6.7.5, which was acting on a rank-2 tensor, this one acts on a rank-3 tensor $(\delta\Gamma^{\rho}_{\alpha\nu})$. That means it has three Christoffel terms rather than just two:

$$\nabla_{\mu}T^{\rho}_{\alpha\nu} = \frac{\partial T^{\rho}_{\alpha\nu}}{\partial x^{\mu}} + \Gamma^{\rho}_{\mu\lambda}T^{\lambda}_{\alpha\nu} - \Gamma^{\lambda}_{\mu\alpha}T^{\rho}_{\lambda\nu} - \Gamma^{\lambda}_{\mu\nu}T^{\rho}_{\lambda\alpha}, \tag{8.3.10}$$

a positive one for the contravariant index and a negative one for each of the covariant indices. If you're getting caught up with the indices, just remember Christoffel symbols are symmetric in the bottom two (i.e. $\Gamma^{\rho}_{\alpha\nu} = \Gamma^{\rho}_{\nu\alpha}$). As a result of this observation, we can say

$$\delta R^{\rho}_{\alpha\mu\nu} = \nabla_{\mu}\left(\delta\Gamma^{\rho}_{\alpha\nu}\right) - \nabla_{\nu}\left(\delta\Gamma^{\rho}_{\alpha\mu}\right), \tag{8.3.11}$$

which is a little easier to look at and is going to be more useful later.

Now that we have a simple representation for the variation of the Riemann tensor, we can contract to acquire the variation in the Ricci tensor. This results in

$$\delta R_{\alpha\nu} = \delta R^{\rho}_{\alpha\rho\nu} = \nabla_{\rho}\left(\delta\Gamma^{\rho}_{\alpha\nu}\right) - \nabla_{\nu}\left(\delta\Gamma^{\rho}_{\alpha\rho}\right), \tag{8.3.12}$$

where ρ has become a summation index. The original term we need to make vanish is

$$g^{\alpha\nu}\frac{\delta R_{\alpha\nu}}{\delta g^{\alpha\nu}} = \frac{g^{\alpha\nu}}{\delta g^{\alpha\nu}}\left[\nabla_{\rho}\left(\delta\Gamma^{\rho}_{\alpha\nu}\right) - \nabla_{\nu}\left(\delta\Gamma^{\rho}_{\alpha\rho}\right)\right],$$

but we'll need to move back a little further in our work to see this happen. This was a originally a term inside an integral (Recall Eq. 8.3.3). We also pulled out a $\delta g^{\alpha\nu}$ and canceled a $\sqrt{|\det(g)|}$ along the way, so

$$\int g^{\alpha\nu}\frac{\delta R_{\alpha\nu}}{\delta g^{\alpha\nu}}\delta g^{\alpha\nu}\sqrt{|\det(g)|}\,d^4x = \int g^{\alpha\nu}\delta R_{\alpha\nu}\sqrt{|\det(g)|}\,d^4x$$

is what actually vanishes. Using Eq. 8.3.12, this terms is

$$\int g^{\alpha\nu}\left[\nabla_\rho\left(\delta\Gamma^\rho_{\alpha\nu}\right) - \nabla_\nu\left(\delta\Gamma^\rho_{\alpha\rho}\right)\right]\sqrt{|\det(g)|}\,d^4x,$$

but it still needs just a little more work. We can distribute the $g^{\alpha\nu}$ to get

$$\int \left[g^{\alpha\nu}\nabla_\rho\left(\delta\Gamma^\rho_{\alpha\nu}\right) - g^{\alpha\nu}\nabla_\nu\left(\delta\Gamma^\rho_{\alpha\rho}\right)\right]\sqrt{|\det(g)|}\,d^4x.$$

Since the symbol used for summation indices is meaningless, we can say $\nabla_\rho\left(\delta\Gamma^\rho_{\alpha\nu}\right) = \nabla_\lambda\left(\delta\Gamma^\lambda_{\alpha\nu}\right)$ and $g^{\alpha\nu}\nabla_\nu = g^{\alpha\lambda}\nabla_\lambda$. This gives

$$\int \left[g^{\alpha\nu}\nabla_\lambda\left(\delta\Gamma^\lambda_{\alpha\nu}\right) - g^{\alpha\lambda}\nabla_\lambda\left(\delta\Gamma^\rho_{\alpha\rho}\right)\right]\sqrt{|\det(g)|}\,d^4x.$$

We also know the covariant derivative of the metric is always zero ($\nabla_\lambda g^{\alpha\nu} = 0$), so we can pull out the covariant derivative arriving at

$$\int \nabla_\lambda\left[g^{\alpha\nu}\delta\Gamma^\lambda_{\alpha\nu} - g^{\alpha\lambda}\delta\Gamma^\rho_{\alpha\rho}\right]\sqrt{|\det(g)|}\,d^4x \tag{8.3.13}$$

Now we're talking!

What we have now is the covariant derivative integrated over the *entire* 4-D "volume" of spacetime. Remember the curl theorem (Eq. 3.5.12) from vector calculus? That was in 3-D space, but it does generalize to higher dimension tensors and its distinction with the divergence theorem blurs a bit. This is typically written as

$$\int_{\text{whole}} dT = \int_{\text{boundary}} T, \tag{8.3.14}$$

but that's a bit general for my taste. Essentially, it says the rate of some tensor T integrated (i.e. infinitesimally summed) over a whole space is equal to

the tensor T integrated (i.e. infinitesimally summed) over the space's boundary. When applied to Eq. 8.3.13, this tells us we can just sum the contributions of

$$g^{\alpha\nu}\delta\Gamma^{\lambda}_{\alpha\nu} - g^{\alpha\lambda}\delta\Gamma^{\rho}_{\alpha\rho}$$

over the boundary of all spacetime (i.e. infinity). It is common to assume spacetime is flat when we're infinitely far from the source of gravity. If we do this here, the coordinates become simple curvilinear coordinates, which don't vary much at infinity. This means $\delta\Gamma^{\lambda}_{\alpha\nu} = 0$ and the entire term vanishes as desired.

8.4 Sweating the Details

Now that we've seen two different approaches for deriving Einstein's equation (Eq. 8.2.6), we need to make sense of it. So far we only know that matter can bend (or warp) space, but deep understanding is in the details. Let's start by examining our new representation of the matter.

Stress-Energy Tensor

It turns out that matter just isn't enough to describe what occupies (and affects) a space. If we recall that $E_p = m_p c^2$ means that mass is just a type of energy, then it becomes clear we need to consider *all* the energy occupying a space. This is where the **stress-energy tensor** comes in because it includes so much more than just mass. We usually work with it in contravariant form:

$$T^{\alpha\nu} \longrightarrow \begin{bmatrix} \mathcal{E} & \Phi_1 & \Phi_2 & \Phi_3 \\ \Phi_1 & P_1 & \sigma_{12} & \sigma_{13} \\ \Phi_2 & \sigma_{21} & P_2 & \sigma_{23} \\ \Phi_3 & \sigma_{31} & \sigma_{32} & P_3 \end{bmatrix}, \tag{8.4.1}$$

where $\mathcal{E}$ is energy density, P is pressure (i.e. compressive or tensile stress), and σ is shear stress. The vector $[\Phi_1, \Phi_2, \Phi_3]$ is energy flux or, equivalently, momentum density (by symmetry $T^{\alpha\nu} = T^{\nu\alpha}$).

The energy density is just the energy per unit volume, so it simply represents the position of the energy. The stress and pressure components tell us how portions of that energy are affecting other portions. Finally, the energy

flux (or momentum density) tells us how the energy is moving. As a result, more than just the energy's existence, its interactions and motion can also affect the curvature of spacetime. Another way to think about this is it's both potential energy *and* kinetic energy that curve spacetime.

This tensor obeys a form of the principle of conservation of energy-momentum (i.e. 4-momentum, see Eq. 7.4.23):

$$\nabla_\nu T^{\alpha\nu} = 0\,, \tag{8.4.2}$$

where ν is a summation index. It's important to note the stress-energy tensor is defined at a single position in spacetime (i.e. an event), so it is a function of both space and time in general. It is also zero where there is no energy (i.e. anywhere in the vacuum of spacetime).

Some Context

A massive body like our Sun can be said to hold onto all the planets, asteroids, comets, etc. simply with energy density. That component of Einstein's equation (Eq. 8.2.6), namely

$$R_{tt} - \frac{1}{2} g_{tt} R = \frac{8\pi G}{c^4} T_{tt},$$

simplifies to Eq. 8.2.1 in the **weak-field approximation**. Yes, I'm saying the Sun creates a weak field. For comparison, a strong field is created by something like a super-giant star or a black hole. Our sun isn't called a yellow *dwarf* for nothing. However, the orbit of Mercury *noticeably* wobbles being so close to the Sun, which was a phenomenon we were unable to explain until general relativity. From a practical point of view, we really only need Einstein's equation (Eq. 8.2.6) when classical physics isn't enough.

Let's consider something a little more exciting: a black hole. Black holes (i.e. objects so massive that not even light can escape) had been speculated for over a century before the publication of general relativity. However, the term "black hole" wasn't coined until physicist John Wheeler first used it in the 1970s. Understanding black holes requires all the components in the stress-energy tensor (Eq. 8.4.1). They curve spacetime by not only existing, but also traveling through space, rotating, and forming orbits with stars and other black holes. All of these motions affect spacetime in different ways. Rotation can twist spacetime into a spiral and it's even speculated

that wobbles can create waves in spacetime. There's also a bit of lag since all these the effects only propagate at the speed of light.

Weird Units

Some of the components of the stress-energy tensor (Eq. 8.4.1) seem to have some units that don't match, but they do if we're careful. Energy density has units of $\mathrm{J/m^3}$ in the SI system, so we'll use that as a reference. Pressure and stress have a unit of $\mathrm{N/m^2}$, but we get

$$\frac{\mathrm{N}}{\mathrm{m^2}} = \frac{\mathrm{N\ m}}{\mathrm{m^3}} = \frac{\mathrm{J}}{\mathrm{m^3}}$$

with a little manipulation. Energy flux the rate at which energy passes through a surface (called "intensity" with regard to waves) and has units of $\mathrm{W/m^2}$. With a little manipulation, this becomes

$$\frac{\mathrm{W}}{\mathrm{m^2}} = \frac{\mathrm{J}}{\mathrm{s\ m^2}} = \frac{\mathrm{J}}{\mathrm{m^3}}\frac{\mathrm{m}}{\mathrm{s}},$$

which varies from the expected unit by m/s. This turns out to be just a factor of $c = 3\times10^8$ m/s. A similar unit phenomenon happens to momentum density with a unit of

$$\frac{\mathrm{kg\ m/s}}{\mathrm{m^3}} = \frac{\mathrm{kg\ m^2/s^2}}{\mathrm{m^3}}\frac{\mathrm{s}}{\mathrm{m}} = \frac{\mathrm{J}}{\mathrm{m^3}}\frac{\mathrm{s}}{\mathrm{m}},$$

which varies from the expected unit by s/m (i.e. a factor of $1/c$).

Recall for Eq. 7.3.6, we introduced a notation changing the contravariant coordinates from (ct, x, y, z) to (x^0, x^1, x^2, x^3). Specifically, this states $x^0 \equiv ct$, which means we'd be measuring time in spatial units (e.g. meters). I know this seems weird, but spacetime fails to distinguish between space and time, so it's actually more physical to do the same on paper. As a result of this, the speed of light becomes

$$c = 299,792,458\ \frac{\mathrm{m}}{\mathrm{s}} = 1\ \frac{\mathrm{m}}{\mathrm{m}} = 1,$$

so the unit of the stress-energy tensor (Eq. 8.4.1) becomes $\mathrm{J/m^3}$ as expected for all components. The quantity c is now simply a unit conversion between meters and seconds. We actually did this without realizing it throughout

Chapter 7 with the use of $\beta = v/c$ (e.g. half the speed is light was simply $\beta = 0.5$). The only difference now is that we're openly embracing it.

Traditionally, proponents of general relativity have gone a step further. Since the quantity $G = 6.674 \times 10^{-11}\ \mathrm{Nm^2/kg^2}$ is in Einstein's equation (Eq. 8.2.6), it shows up quite often. Physicist get a bit lazy sometimes and stop writing it. In other words, they set

$$G = 6.67408 \times 10^{-11}\ \frac{\mathrm{Nm^2}}{\mathrm{kg^2}} = 1,$$

so that all the G's disappear. Ok, so maybe it's not just laziness. Theoretical physicists tend to be unconcerned with universal constants since they don't actually say much about the relationship itself. Their only purpose to make the relationships match experiment. What I'm saying is this isn't really a new thing to set a constant to one. It's referred to as **natural units**.

The consequence of setting both $c = 1$ and $G = 1$ is called **geometrized units** because the units of all the quantities relevant to general relativity reduce to variations of only the meter, the unit of geometry. We end up with unit conversions like

$$\begin{cases} \dfrac{G}{c^2} = 7.42592 \times 10^{-28}\ \dfrac{\mathrm{m}}{\mathrm{kg}}; & \text{mass} \\ \dfrac{G}{c^3} = 2.47702 \times 10^{-36}\ \dfrac{\mathrm{m}}{\mathrm{Ns}}; & \text{linear and angular momentum} \\ \dfrac{G}{c^4} = 8.26245 \times 10^{-45}\ \dfrac{1}{\mathrm{N}}; & \text{force, energy, energy density, pressure} \\ \dfrac{G}{c^5} = 2.75606 \times 10^{-53}\ \dfrac{1}{\mathrm{W}}; & \text{power} \end{cases} \tag{8.4.3}$$

and the size of these conversions drastically brings the large astronomical values down to comprehensible ones. For example, the mass of the sun is now

$$M_\odot = 1.989 \times 10^{30}\ \mathrm{kg} \left(7.42592 \times 10^{-28}\ \frac{\mathrm{m}}{\mathrm{kg}}\right) = 1477\ \mathrm{m},$$

which really makes no conceptual sense whatsoever. However, with less to carry through the math, there is less chance of calculation error.

As you can see in Table 8.1, all the quantities in the stress-energy tensor now have a unit of $1/\mathrm{m^2}$ and we no longer have to worry about the discrepancy. Furthermore, these new units change all the equations we use as well.

Table 8.1: This is a list of quantities relevant to general relativity and their corresponding geometrized unit.

Quantity	Geometrized Unit
Length	m
Time	m
Mass	m
Energy	m
Linear Momentum	m
Angular Momentum	m^2
Energy Density	$1/\mathrm{m}^2 = \mathrm{m}^{-2}$
Momentum Density	$1/\mathrm{m}^2 = \mathrm{m}^{-2}$
Energy Flux	$1/\mathrm{m}^2 = \mathrm{m}^{-2}$
Pressure	$1/\mathrm{m}^2 = \mathrm{m}^{-2}$
Stress	$1/\mathrm{m}^2 = \mathrm{m}^{-2}$
Force	unitless
Power	unitless

For example, Einstein's equation (Eq. 8.2.6) reduces to

$$R_{\alpha\nu} - \frac{1}{2} g_{\alpha\nu} R = 8\pi \, T_{\alpha\nu} \, . \tag{8.4.4}$$

If you're still having trouble conceptualizing when you're done working through the math, then you can always convert the final result back to SI units to interpret it.

8.5 Special Cases

Throughout the last few sections, we've been dealing with general relativity without applying it to anything specific. It was important to get some groundwork laid first. I'd like to take a little time in this section to briefly mentioned some specific contexts where Einstein's equation (Eq. 8.4.4) can be and is often applied. We'll also be working out some details through example.


Karl Schwarzschild


Georges Lemaître

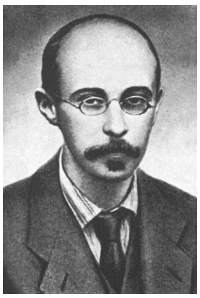

Alexander Friedmann

Figure 8.4: These people were important in the application of general relativity.

Spherical Symmetry

It is very common for large objects like stars to be **spherically symmetric**, which just means there is no angular dependence within the star. Only changes in radial distance from the center result in changes in the star's properties. Furthermore, most stars tend to rotate slowly (e.g. the Sun takes about a month to make one full rotation), so it's safe to assume the star is also **static** (i.e. has temporal symmetry). This means its properties don't change in time.

If the star is spherically symmetric, then it's angular terms should be identical to the standard spherical metric terms (Eq. 7.2.6). If the star is also static, then none of its terms should should be functions of time. Therefore, the metric tensor takes the form:

$$g_{\alpha\delta} \longrightarrow \begin{bmatrix} -a(r) & 0 & 0 & 0 \\ 0 & b(r) & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2\theta \end{bmatrix}, \tag{8.5.1}$$

where a and b are arbitrary functions of radial distance. Using Eq. 7.2.3, the line element takes the form:

$$ds^2 = -a(r)\,dt^2 + b(r)\,dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2 , \tag{8.5.2}$$

for both *inside* and *outside* a spherically symmetric (and static) star.

Example 8.5.1

Show that the metric for spherically symmetric (and static) star is diagonal.

- Mathematically speaking, spherically symmetry this tells us swapping angular variables, $\theta \to -\theta$ and/or $\phi \to -\phi$, gives the same result. Also, the star having temporal symmetry means $t \to -t$ gives the same result. These are all coordinate transformations and we know from Section 6.6 that all covariant tensors (e.g. $g_{\alpha\delta}$) transform by Eq. 6.6.3.

- The time transformation shows that

$$g'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\delta}{\partial x'^\nu} g_{\alpha\delta}$$

$$g'_{\mu t} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\delta}{\partial t'} g_{\alpha\delta}.$$

  If we expand the sum over δ, then

$$g'_{\mu t} = \frac{\partial x^\alpha}{\partial x'^\mu}\left[\frac{\partial t}{\partial t'}g_{\alpha t} + \frac{\partial r}{\partial t'}g_{\alpha r} + \frac{\partial \theta}{\partial t'}g_{\alpha\theta} + \frac{\partial \phi}{\partial t'}g_{\alpha\phi}\right]$$

  and, since the coordinates are orthogonal and $t' = -t$, we get

$$g'_{\mu t} = \frac{\partial x^\alpha}{\partial x'^\mu}\left[(-1)\, g_{\alpha t} + (0)\, g_{\alpha r} + (0)\, g_{\alpha\theta} + (0)\, g_{\alpha\phi}\right] = -\frac{\partial x^\alpha}{\partial x'^\mu} g_{\alpha t}.$$

- Note, α is still a summation index but μ is a free index, which means this is still four separate equations. Expanding over the final sum, we get

$$g'_{\mu t} = -\left[\frac{\partial t}{\partial x'^\mu}g_{tt} + \frac{\partial r}{\partial x'^\mu}g_{rt} + \frac{\partial \theta}{\partial x'^\mu}g_{\theta t} + \frac{\partial \phi}{\partial x'^\mu}g_{\phi t}\right],$$

  which is still four equations due to the free index μ. For $\mu = t$, this is

$$g'_{tt} = -\left[\frac{\partial t}{\partial t'}g_{tt} + \frac{\partial r}{\partial t'}g_{rt} + \frac{\partial \theta}{\partial t'}g_{\theta t} + \frac{\partial \phi}{\partial t'}g_{\phi t}\right],$$

$$g'_{tt} = -\left[(-1)\, g_{tt} + (0)\, g_{rt} + (0)\, g_{\theta t} + (0)\, g_{\phi t}\right] = +g_{tt},$$

which shows it's unchanged under the transformation. However, for $\mu = r$, this is

$$g'_{rt} = -\left[\frac{\partial t}{\partial r'}g_{tt} + \frac{\partial r}{\partial r'}g_{rt} + \frac{\partial \theta}{\partial r'}g_{\theta t} + \frac{\partial \phi}{\partial r'}g_{\phi t}\right],$$

$$g'_{rt} = -\left[(0)\, g_{tt} + (+1)\, g_{rt} + (0)\, g_{\theta t} + (0)\, g_{\phi t}\right] = -g_{rt},$$

which is a problem. If the star has temporal symmetry, then $g'_{rt} = g_{rt}$ so we must conclude that $g_{rt} = 0$. In the same way, $g_{\theta t} = 0$ and $g_{\phi t} = 0$.

- We can perform this same process on the spherical symmetry transformations, $\theta \to -\theta$ and/or $\phi \to -\phi$. Including the work for it here would be redundant since all we'd be changing would be indices. The results are as follows:

$$g'_{\theta\theta} = g_{\theta\theta} \text{ and } g'_{\phi\phi} = g_{\phi\phi},$$

implying these can be non-zero like g_{tt}, and all off-diagonal terms are zero. You can save yourself a little time knowing that the metric tensor is always symmetric (ie. $g_{\alpha\delta} = g_{\delta\alpha}$).

Example 8.5.2

Determine the Christoffel symbols and curvature tensors in the space occupied by a spherically symmetric (and static) star where the metric is given by Eq. 8.5.1.

- There are quite a few components in these quantities and the process gets a bit repetitive. I'll save time by deriving only one of each. You can find an entire list of curvatures for a variety of geometries in Appendix C.

- Christoffel symbols can be found using Eq. 6.7.6. We've done this for an arbitrary 3-space in Example 6.7.1, but this generalizes to 4-space with

$$\Gamma^{\delta}_{\mu\nu} = \frac{1}{2}g^{\lambda\delta}\left(\frac{\partial g_{\lambda\mu}}{\partial x^{\nu}} + \frac{\partial g_{\lambda\nu}}{\partial x^{\mu}} - \frac{\partial g_{\mu\nu}}{\partial x^{\lambda}}\right),$$

where λ is a summation index. For the spherically symmetric geometry given, we'll perform the steps for

$$\Gamma^t_{tr} = \frac{1}{2} g^{\mu t} \left(\frac{\partial g_{\mu t}}{\partial r} + \frac{\partial g_{\mu r}}{\partial t} - \frac{\partial g_{tr}}{\partial x^\mu} \right),$$

where λ is a summation index. Since the inverse metric is diagonal, the only non-zero terms occur when $\mu = t$ because of the $g^{\mu t}$ out front. The result is

$$\Gamma^t_{tr} = \frac{1}{2} g^{tt} \left(\frac{\partial g_{tt}}{\partial r} + \frac{\partial g_{tr}}{\partial t} - \frac{\partial g_{tr}}{\partial t} \right)$$

$$\Gamma^t_{tr} = \frac{1}{2} g^{tt} \frac{\partial g_{tt}}{\partial r}$$

Since the metric tensor is diagonal, we know $g^{tt} = 1/g_{tt}$ and we get

$$\Gamma^t_{tr} = \frac{1}{2 g_{tt}} \frac{\partial g_{tt}}{\partial r} = \boxed{\frac{1}{2a} \frac{\partial a}{\partial r}}$$

- Using the Christoffel symbols, we can get the Riemann curvatures. We'll go with R^t_{rtr} for our work. Using Eq. 8.1.1, we get

$$R^t_{rtr} = \frac{\partial \Gamma^t_{rr}}{\partial t} - \frac{\partial \Gamma^t_{rt}}{\partial r} + \Gamma^t_{\lambda t} \Gamma^\lambda_{rr} - \Gamma^t_{\lambda r} \Gamma^\lambda_{rt},$$

where λ is a summation index. Since the summation shows up twice, that's a total of 8 non-derivative terms. However, judging from the non-zero Christoffel symbols in Section C.5, we can say only $\lambda = r$ in the first summation results in a non-zero value and only $\lambda = t$ does in the second. Also, $\Gamma^t_{rr} = 0$, not that it matters since none are functions of time anyway. Therefore,

$$R^t_{rtr} = -\frac{\partial \Gamma^t_{rt}}{\partial r} + \Gamma^t_{rt} \Gamma^r_{rr} - \Gamma^t_{tr} \Gamma^t_{rt}$$

$$R^t_{rtr} = \left[-\frac{\partial}{\partial r} \left(\frac{1}{2a} \frac{\partial a}{\partial r} \right) \right] + \left(\frac{1}{2a} \frac{\partial a}{\partial r} \right) \left(\frac{1}{2b} \frac{\partial b}{\partial r} \right) - \left(\frac{1}{2a} \frac{\partial a}{\partial r} \right) \left(\frac{1}{2a} \frac{\partial a}{\partial r} \right)$$

$$R^t_{rtr} = \left[-\frac{\partial}{\partial r}\left(\frac{1}{2a}\right)\frac{\partial a}{\partial r} - \frac{1}{2a}\frac{\partial}{\partial r}\left(\frac{\partial a}{\partial r}\right)\right] + \frac{1}{4ab}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} - \frac{1}{4a^2}\left(\frac{\partial a}{\partial r}\right)^2$$

$$R^t_{rtr} = \frac{1}{2a^2}\left(\frac{\partial a}{\partial r}\right)^2 - \frac{1}{2a}\frac{\partial^2 a}{\partial r^2} + \frac{1}{4ab}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} - \frac{1}{4a^2}\left(\frac{\partial a}{\partial r}\right)^2$$

$$\boxed{R^t_{rtr} = -\frac{1}{2a}\frac{\partial^2 a}{\partial r^2} + \frac{1}{4ab}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} + \frac{1}{4a^2}\left(\frac{\partial a}{\partial r}\right)^2}$$

- We could repeat this with Eq. 8.1.5 to get the Ricci curvatures. However, if we have all the Riemann curvatures, then it's easier to just contract the Riemann tensor with

  $$R_{\alpha\nu} = R^\mu_{\alpha\mu\nu},$$

  where μ is a summation index. Again, we'll pick just one to solve:

  $$R_{tt} = R^\mu_{t\mu t} = R^t_{ttt} + R^r_{trt} + R^\theta_{t\theta t} + R^\phi_{t\phi t}$$

  Using the Riemann curvatures from Section C.5, we get

  $$R_{tt} = [0] + \left[-\frac{1}{4ab}\left(\frac{\partial a}{\partial r}\right)^2 - \frac{1}{4b^2}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} + \frac{1}{2a}\frac{\partial^2 a}{\partial r^2}\right] + \left[\frac{1}{2rb}\frac{\partial a}{\partial r}\right] + \left[\frac{1}{2rb}\frac{\partial a}{\partial r}\right]$$

  $$\boxed{R_{tt} = -\frac{1}{4ab}\left(\frac{\partial a}{\partial r}\right)^2 - \frac{1}{4b^2}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} + \frac{1}{2a}\frac{\partial^2 a}{\partial r^2} + \frac{1}{rb}\frac{\partial a}{\partial r}}$$

- Using Eq. 8.1.6 (just another contraction), the Ricci curvature scalar is given by

  $$R = g^{\alpha\nu}R_{\alpha\nu} = g^{\alpha t}R_{\alpha t} + g^{\alpha r}R_{\alpha r} + g^{\alpha\theta}R_{\alpha\theta} + g^{\alpha\phi}R_{\alpha\phi}.$$

  Luckily, we know both the metric and the Ricci tensor are diagonal, so

  $$R = g^{tt}R_{tt} + g^{rr}R_{rr} + g^{\theta\theta}R_{\theta\theta} + g^{\phi\phi}R_{\phi\phi}.$$

  Using the Ricci curvatures from Section C.5 and combining like terms, we get

  $$\boxed{R = \frac{2}{r^2}\left(1 - \frac{1}{b}\right) - \frac{2}{rab}\frac{\partial a}{\partial r} + \frac{1}{2a^2b}\left(\frac{\partial a}{\partial r}\right)^2 + \frac{2}{rb^2}\frac{\partial b}{\partial r} + \frac{1}{2ab^2}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} - \frac{1}{ab}\frac{\partial^2 a}{\partial r^2}}$$

Example 8.5.3

Determine a convenient orthnormal basis for the space occupied by a spherically symmetric (and static) star where the coordinates are given by the metric in Eq. 8.5.1.

- A generalization of Eq. 6.4.9 to four-dimensional spacetime is

$$g_{\hat{\mu}\hat{\delta}} = (\hat{e}_\mu)^\alpha \, (\hat{e}_\delta)^\nu \, g_{\alpha\nu} \longrightarrow \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

  meaning the metric mimics flat spacetime in the orthonormal basis. Since we're building an orthonormal basis *from* an already orthogonal coordinate basis, each orthonormal basis vector will only have one non-zero component in the coordinate basis. This will drastically simplify our summations.

- We'll start with time component of the time vector $(\hat{e}_t)^t$:

$$g_{\hat{t}\hat{t}} = (\hat{e}_t)^\alpha \, (\hat{e}_t)^\nu \, g_{\alpha\nu}.$$

  However, we already know $\alpha = \nu = t$ is the only non-zero component, so

$$g_{\hat{t}\hat{t}} = (\hat{e}_t)^t \, (\hat{e}_t)^t \, g_{tt}$$

$$-1 = \left[(\hat{e}_t)^t\right]^2 g_{tt} \quad \Rightarrow (\hat{e}_t)^t = \frac{1}{\sqrt{-g_{tt}}}$$

- The radial component of the radial vector works out in a similar way as

$$g_{\hat{r}\hat{r}} = (\hat{e}_r)^\alpha \, (\hat{e}_r)^\nu \, g_{\alpha\nu} = (\hat{e}_r)^r \, (\hat{e}_r)^r \, g_{rr}$$

$$+1 = \left[(\hat{e}_r)^r\right]^2 g_{rr} \quad \Rightarrow (\hat{e}_r)^r = \frac{1}{\sqrt{g_{rr}}}.$$

  The angular components of the angular vectors are identical in pattern.

- Therefore, the four orthonormal basis vectors take the form

$$\left\{\begin{array}{rcl} \hat{e}_t &=& \left[\frac{1}{\sqrt{-g_{tt}}}, 0, 0, 0\right] \\ \hat{e}_r &=& \left[0, \frac{1}{\sqrt{g_{rr}}}, 0, 0\right] \\ \hat{e}_\theta &=& \left[0, 0, \frac{1}{\sqrt{g_{\theta\theta}}}, 0\right] \\ \hat{e}_\phi &=& \left[0, 0, 0, \frac{1}{\sqrt{g_{\phi\phi}}}\right] \end{array}\right\} \tag{8.5.3}$$

and, using Eq. 8.5.1, we get

$$\boxed{\left\{\begin{array}{rcl} \hat{e}_t &=& \left[\frac{1}{\sqrt{a}}, 0, 0, 0\right] \\ \hat{e}_r &=& \left[0, \frac{1}{\sqrt{b}}, 0, 0\right] \\ \hat{e}_\theta &=& \left[0, 0, \frac{1}{r}, 0\right] \\ \hat{e}_\phi &=& \left[0, 0, 0, \frac{1}{r\sin\theta}\right] \end{array}\right\}} \tag{8.5.4}$$

Eq. 8.5.2 is nice and simple, but it has it's limitations. It assumes the star *never* changes. Eventually, every star, rotating or not, is going to collapse. As long as your star maintains spherical symmetry perfectly during the collapse, then you can say

$$ds^2 = -a(t,r)\,dt^2 + b(t,r)\,dr^2 + r^2 d\theta^2 + r^2\sin^2\theta\,d\phi^2, \tag{8.5.5}$$

where a and b are now arbitrary functions of both radial distance and time. You just have to be careful about the conditions of your star's collapse.

Perfect Fluids

A star happens to be made of plasma, but plasma behave very similarly to fluids. If our star is not very viscous, free of shear stress, and has only isotropic pressures (i.e. the pressure is independent of direction); then we call it a **perfect fluid**. This is common for a spherically symmetric star. Under these conditions, the stress-energy tensor (Eq. 8.4.1) takes the form:

$$T^{\alpha\nu} = (\rho + P)u^\alpha u^\nu + g^{\alpha\nu}P \tag{8.5.6}$$

where u^α is the 4-velocity (Eq. 7.4.3), $\rho(r)$ is the density at r, and $P(r)$ is the pressure at r. See Section 8.4 for a more general description of the stress-energy tensor.

If we're dealing with a star that is also static, then the fluid is not moving in space (only through time). That means its 4-velocity is

$$u^\alpha \longrightarrow \begin{bmatrix} \frac{1}{\sqrt{a}} \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{8.5.7}$$

where $a(r)$ is from the spherically symmetric line element (Eq. 8.5.2). The $1/\sqrt{a}$ is due to a scale factor we picked up since we're working in a *coordinate* basis rather than an *orthonormal* basis (see Example 8.5.3 for more details). In other words, using Eq. 6.4.8, the components of the 4-velocity are

$$u^\alpha = (\hat{e}_\lambda)^\alpha \, u^{\hat{\lambda}},$$

where the orthonormal basis vectors are given in Eq. 8.5.3. The result is $u^t = 1/\sqrt{a}$, but $u^{\hat{t}} = 1$. It gets really weird, so I do everything I can to stick with the coordinate basis in general relativity. If we plug Eq. 8.5.7 into Eq. 8.5.6, then we get

$$\left\{ \begin{array}{rcl} T^{tt} &=& (\rho + P) u^t u^t + g^{tt} P \\ T^{rr} &=& g^{rr} P \\ T^{\theta\theta} &=& g^{\theta\theta} P \\ T^{\phi\phi} &=& g^{\phi\phi} P \end{array} \right\}$$

$$\left\{ \begin{array}{rcl} T^{tt} &=& \rho/a \\ T^{rr} &=& P/b \\ T^{\theta\theta} &=& P/r^2 \\ T^{\phi\phi} &=& \dfrac{P}{r^2 \sin^2\theta} \end{array} \right\} \tag{8.5.8}$$

for the four non-zero components of the stress-energy tensor for a perfect static fluid.

Example 8.5.4

Determine the form of $b(r)$ in Eq. 8.5.1 for a spherically symmetric star composed of perfect static fluid.

- We'll start with Einstein's equation (Eq. 8.4.4), but only the

$$R_{tt} - \frac{1}{2} g_{tt} R = 8\pi\, T_{tt}, \tag{8.5.9}$$

  component is necessary.

- Unfortunately, the stress-energy tensor in Eq. 8.5.8 is contravariant and we need it to be covariant. We can use the metric tensor to bring down both indices with

$$T_{tt} = g_{t\alpha} g_{t\nu} T^{\alpha\nu}.$$

  That has 16 terms, but since the metric tensor is diagonal, we know $\alpha = \nu = t$ leaving us with just one non-zero term:

$$T_{tt} = g_{tt} g_{tt} T^{tt} = (-a)\,(-a)\left(\frac{\rho}{a}\right) = a\rho.$$

- From Section C.5, we know

$$R_{tt} = \frac{1}{rb}\frac{\partial a}{\partial r} - \frac{1}{4ab}\left(\frac{\partial a}{\partial r}\right)^2 - \frac{1}{4b^2}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} + \frac{1}{2b}\frac{\partial^2 a}{\partial r^2}$$

  and

$$R = \frac{2}{r^2}\left(1 - \frac{1}{b}\right) - \frac{2}{rab}\frac{\partial a}{\partial r} + \frac{1}{2a^2 b}\left(\frac{\partial a}{\partial r}\right)^2 + \frac{2}{rb^2}\frac{\partial b}{\partial r} + \frac{1}{2ab^2}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} - \frac{1}{ab}\frac{\partial^2 a}{\partial r^2},$$

  which are also part of Eq. 8.5.9 along with the metric tensor.

- Substituting all these into Eq. 8.5.9 and combining like terms results in

$$\frac{a}{r^2}\left(1 - \frac{1}{b}\right) + \frac{a}{rb^2}\frac{\partial b}{\partial r} = 8\pi\, a\rho.$$

  If we multiply through by r^2/a and factor out a 2 on the right, we get

$$\left(1 - \frac{1}{b}\right) + \frac{r}{b^2}\frac{\partial b}{\partial r} = 2\left(4\pi\, r^2 \rho\right).$$

Since $\partial r/\partial r = 1$ and

$$\frac{1}{b^2}\frac{\partial b}{\partial r} = \frac{\partial}{\partial r}\left(1 - \frac{1}{b}\right),$$

we can say

$$\left(1 - \frac{1}{b}\right)\frac{\partial r}{\partial r} + r\frac{\partial}{\partial r}\left(1 - \frac{1}{b}\right) = 2\left(4\pi\, r^2\rho\right).$$

By doing the derivative chain rule (Eq. 3.1.2) in reverse, this is

$$\frac{\partial}{\partial r}\left[r\left(1 - \frac{1}{b}\right)\right] = 2\left(4\pi\, r^2\rho\right).$$

and integrating both sides over r gives us

$$r\left(1 - \frac{1}{b}\right) = 2\int_0^r \left(4\pi\, r^2\rho\right) dr.$$

- It might appear we're at a stand still, but the integral on the left is something special. The mass enclosed in a sphere of radius r (centered at the center of the star) is given by

$$m(r) = \int_0^{2\pi}\int_0^{\pi}\int_0^r \rho(r)\; r^2 \sin\theta\, dr\, d\theta\, d\phi,$$

but evaluating the θ and ϕ integrals simplifies this to

$$m(r) = 4\pi \int_0^r \rho(r)\; r^2 dr,$$

which is exactly our integral on the right. That means we get

$$r\left(1 - \frac{1}{b}\right) = 2m$$

and solving for b gives us

$$1 - \frac{1}{b} = \frac{2m}{r} \quad \Rightarrow \frac{1}{b} = 1 - \frac{2m}{r}$$

$$\Rightarrow b = \left(1 - \frac{2m}{r}\right)^{-1}.$$

Writing this a little more clearly, we have

$$\boxed{b(r) = \left(1 - \frac{2m(r)}{r}\right)^{-1}}, \tag{8.5.10}$$

where $m(r)$ is the mass enclosed by a sphere of radius r (centered at the center of the star).

The Vacuum

If we limit ourselves to the spacetime *outside* a star, then we're in the vacuum. This is particularly important if we want to know how the star is affecting other objects (e.g. planets, comets, people, etc.). We've mentioned the vacuum in the book before and even used it in Section 5.5 to derive the equations describing electromagnetic waves. A vacuum is just a place (and time) devoid of matter and energy (i.e. empty spacetime). In the case of general relativity, we can say $T_{\alpha\nu} = 0$ anywhere in the vacuum.

Recall, we said Einstein's equation and all quantities in it are defined at a specific event. What we mean is that it doesn't matter if there is a star nearby because $T_{\alpha\nu}$ only has a value for all events *inside* the star. This has consequences for the other quantities in Einstein's equation (Eq. 8.4.4). Substituting in $T_{\alpha\nu} = 0$, we get

$$R_{\alpha\nu} - \frac{1}{2} g_{\alpha\nu} R = 0.$$

There are two ways this equation can be zero: either $R_{\alpha\nu} = 0$ or

$$R_{\alpha\nu} = \frac{1}{2} g_{\alpha\nu} R.$$

However, playing a little with the second possibility gives us

$$g^{\alpha\nu} R_{\alpha\nu} = g^{\alpha\nu} \left(\frac{1}{2} g_{\alpha\nu} R\right)$$

$$g^{\alpha\nu}R_{\alpha\nu} = \frac{1}{2}g^{\alpha\nu}g_{\alpha\nu}R.$$

Since $g^{\alpha\nu}g_{\alpha\nu} = \delta^{\nu}_{\nu} = 4$ (using the Kronecker delta) and Eq. 8.1.6 says $g^{\alpha\nu}R_{\alpha\nu} = R^{\nu}_{\nu} = R$, we ultimately get

$$R = \frac{1}{2}(4R) \quad \Rightarrow 1 = 2,$$

which proves by contradiction that this isn't *really* a possibility. Therefore, we can conclude that

$$R_{\alpha\nu} = 0 \tag{8.5.11}$$

for any α and ν in the vacuum.

Eq. 8.5.2 represents the line element both inside and outside a spherically symmetric (and static) star. If we're limiting ourselves to only *outside* the star, then

$$ds^2 = -\left(1 - \frac{2M}{r}\right)dt^2 + \left(1 - \frac{2M}{r}\right)^{-1}dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2, \tag{8.5.12}$$

where we've replaced $a(r)$ and $b(r)$ with specific functions. This is called the **Schwarzchild solution** since Karl Schwarzschild derived it very shortly after Einstein's publication of general relativity. It is the most famous of the "vacuum solutions" and, by solutions, we mean solutions to Einstein's equation. All *physical* line elements are solutions to Einstein's equation.

Example 8.5.5

Use the Ricci curvatures for spherically symmetric (and static) star found in Section C.5 to derive the Schwarzchild line element (Eq. 8.5.12).

- To solve for the line element, we just need to find the specific forms of $a(r)$ and $b(r)$. We're going to do this using the vacuum condition Eq. 8.5.11, but we have to do it for at least three of the Ricci curvatures to have a solvable system of partial differential equations. Those three

are

$$\left\{\begin{array}{ll} R_{tt}=0= & \frac{1}{rb}\frac{\partial a}{\partial r}-\frac{1}{4ab}\left(\frac{\partial a}{\partial r}\right)^2-\frac{1}{4b^2}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r}+\frac{1}{2b}\frac{\partial^2 a}{\partial r^2} \\ R_{rr}=0= & \frac{1}{4a^2}\left(\frac{\partial a}{\partial r}\right)^2+\frac{1}{rb}\frac{\partial b}{\partial r}+\frac{1}{4ab}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r}-\frac{1}{2a}\frac{\partial^2 a}{\partial r^2} \\ R_{\theta\theta}=0= & 1-\frac{1}{b}-\frac{r}{2ab}\frac{\partial a}{\partial r}+\frac{r}{2b^2}\frac{\partial b}{\partial r} \end{array}\right\}$$

and the $R_{\phi\phi}$ is unnecessary because it's just $R_{\theta\theta}\sin^2\theta$.

- From here on out, this is just a math problem. We can clear all the fractions getting

$$\left\{\begin{array}{ll} 0= & 4ab\frac{\partial a}{\partial r}-br\left(\frac{\partial a}{\partial r}\right)^2-ar\frac{\partial a}{\partial r}\frac{\partial b}{\partial r}+2abr\frac{\partial^2 a}{\partial r^2} \\ 0= & br\left(\frac{\partial a}{\partial r}\right)^2+4a^2\frac{\partial b}{\partial r}+ar\frac{\partial a}{\partial r}\frac{\partial b}{\partial r}-2abr\frac{\partial^2 a}{\partial r^2} \\ 0= & 2ab^2-2ab-br\frac{\partial a}{\partial r}+ar\frac{\partial b}{\partial r} \end{array}\right\}$$

by multiplying by $4ab^2r$, $4a^2br$, and $2ab^2$, respectively.

- Adding the first two equations, several terms cancel and we're left with

$$0=4ab\frac{\partial a}{\partial r}+4a^2\frac{\partial b}{\partial r}$$

$$0=b\frac{\partial a}{\partial r}+a\frac{\partial b}{\partial r}$$

$$0=\frac{\partial}{\partial r}(ab) \quad \Rightarrow ab=\text{constant}$$

or, equivalently, $a=k_1/b$ where k_1 is *not* a function of r (i.e. a constant in the integral over r). Substituting this into the third equation gives us

$$0=2\left(\frac{k_1}{b}\right)b^2-2\left(\frac{k_1}{b}\right)b-br\frac{\partial}{\partial r}\left(\frac{k_1}{b}\right)+\left(\frac{k_1}{b}\right)r\frac{\partial b}{\partial r}$$

$$0 = 2\left(\frac{k_1}{b}\right)b^2 - 2\left(\frac{k_1}{b}\right)b + br\left(\frac{k_1}{b^2}\right)\frac{\partial b}{\partial r} + \left(\frac{k_1}{b}\right)r\frac{\partial b}{\partial r}$$

$$0 = 2b - 2 + \frac{r}{b}\frac{\partial b}{\partial r} + \frac{r}{b}\frac{\partial b}{\partial r}.$$

Combining like terms and clearing fractions by multiplying by $b/2$, we get

$$0 = b^2 - b + r\frac{\partial b}{\partial r}.$$

- Since b is only a function of r, this is just a first-order differential equations we can solve by separation of variables. Rewriting, we get

$$r\frac{db}{dr} = -b\,(b-1) \quad \Rightarrow \frac{-1}{b\,(b-1)}db = \frac{1}{r}dr$$

$$\Rightarrow \left(\frac{1}{b} - \frac{1}{b-1}\right)db = \frac{1}{r}dr.$$

Now we can integrate to get

$$\int\left(\frac{1}{b} - \frac{1}{b-1}\right)db = \int\frac{1}{r}dr$$

$$\ln(b) - \ln(b-1) = \ln(r) - \ln(k_2),$$

where k_2 is *not* a function of r (i.e. a constant in the integral over r). We can use log rules to combine terms and we get

$$\ln\left(\frac{b}{b-1}\right) = \ln\left(\frac{r}{k_2}\right) \quad \Rightarrow \frac{b}{b-1} = \frac{r}{k_2}$$

$$\Rightarrow \frac{k_2}{r}b = b - 1 \quad \Rightarrow 1 = b\left(1 - \frac{k_2}{r}\right),$$

which means

$$\boxed{b = \left(1 - \frac{k_2}{r}\right)^{-1}}$$

and

$$\boxed{a = \frac{k_1}{b} = k_1\left(1 - \frac{k_2}{r}\right)}.$$

- So now we have the general form of both $a(r)$ and $b(r)$. We just need to figure out what k_1 and k_2 look like. We know, as $r \to \infty$, the line element should approach that of flat spacetime (i.e. $a \to -1$). If we take the limit, then

$$-1 = \lim_{r\to\infty} a = \lim_{r\to\infty} k_1\left(1 - \frac{k_2}{r}\right) = k_1,$$

so $\boxed{k_1 = -1}$ and the Schwarzchild solution takes the form

$$ds^2 = -\left(1 - \frac{k_2}{r}\right)dt^2 + \left(1 - \frac{k_2}{r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2. \quad (8.5.13)$$

- Well, k_2 is a little trickier. We know that as the mass of the star approaches zero, then we should also get flat spacetime. When $k_2 \to 0$ we get flat spacetime, but that only tells us that $k_2 \propto M$. We could compare the $b(r)$ here with Eq. 8.5.10 from Example 8.5.4 at the outer boundary of the star. Since $m(r_{\text{star}}) = M$ and the metric should be continuous at the boundary, we get

$$b = \left(1 - \frac{k_2}{r_{\text{star}}}\right)^{-1} = \left(1 - \frac{2M}{r_{\text{star}}}\right)^{-1} \quad \Rightarrow \boxed{k_2 = 2M}.$$

Some of you may be a little uncomfortable with this approach though because Eq. 8.5.10 only applies to a perfect static fluid. For a more rigorous physical approach, see Example 8.6.1.

Example 8.5.6

The time component in the Schwarzchild line element (Eq. 8.5.12) is dependent on r, the distance from the center of the spherically symmetric object. This implies the passage of time is measured differently for observers in different locations in the spacetime curvature. Determine a transformation for time between the following observers:

Observer A: on the Earth's surface,

Observer B: 400 km above the Earth, and

Observer C: very far away from the Earth.

You may ignore the motion of all observers, which is practical assuming the observer A is on the equator and observer B is in geostationary orbit above observer A. Observer C is so far away that the motions of observers A and B don't matter.

- Recall from Section 7.7 that we have to be very careful when discussing who measures what and where they measure it. Since the Schwarzchild line element (Eq. 8.5.12) has no time-dependence, all three observers will have the same *coordinate* time as shown in Figure 8.5. Coordinate time is the time determined by the coordinates we've chosen for the source of curvature (i.e. Earth), which is not something we directly measure. What we measure is our *proper* time and each of the observers has their own because they're all on different world lines.

- Let's assume events 1 and 2 in Figure 8.5 are just two bright flashes of light. These flashes are separated by $\Delta\tau_A$ for the Earth observer. However, those flashes arrive at observer B at events 3 and 4, respectively, separated by $\Delta\tau_B$. Likewise, that's $\Delta\tau_C$ between events 5 and 6 for observer C, the distant observer.

- Assuming none of the observers move through space, Eq. 8.5.12 shows

$$\Delta s_A^2 = -\Delta\tau_A^2 = -\left(1 - \frac{2M}{r_A}\right)\Delta t^2$$

for observer A and

$$\Delta s_B^2 = -\Delta\tau_B^2 = -\left(1 - \frac{2M}{r_B}\right)\Delta t^2$$

for observer B. We can eliminate Δt by dividing these equations, arriving at

$$\frac{\Delta\tau_A^2}{\Delta\tau_B^2} = \frac{1 - 2M/r_A}{1 - 2M/r_B}$$

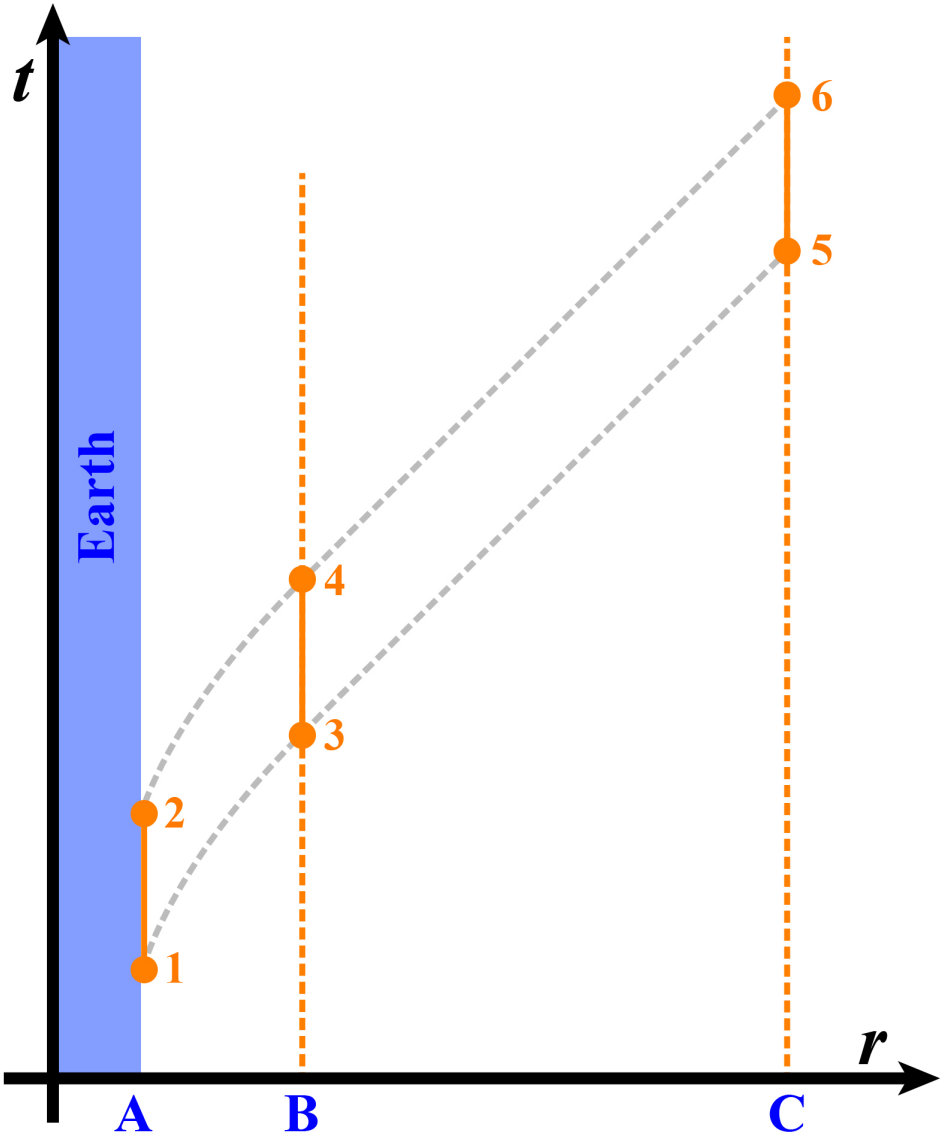


Figure 8.5: Shown here, events 1 and 2 both happen on the Earth's surface. The labels A, B, and C represent the radial distance, r, for each observer in Example 8.5.6. Light travels away from event 1 and 2 along null paths, which are only straight far from the Earth. The curvature has been exaggerated for clarity.

$$\frac{\Delta\tau_A}{\Delta\tau_B} = \sqrt{\frac{1-2M/r_A}{1-2M/r_B}}$$

$$\boxed{\Delta\tau_B = \sqrt{\frac{1-2M/r_B}{1-2M/r_A}}\,\Delta\tau_A} \tag{8.5.14}$$

This shows, as you get closer to the source of gravity (i.e. $r_A < r_B$), time slows down (i.e. $\Delta\tau_A < \Delta\tau_B$). Just be careful! This is in geometrized units (see Table 8.1), so M is measured in meters.

- Observer C is very far away (i.e. $r_C \to \infty$). The light's world line is very straight for them because spacetime is nearly flat. Applying this, Eq. 8.5.14 simplifies to

$$\Delta\tau_C = \frac{1}{\sqrt{1-2M/r_A}}\,\Delta\tau_A.$$

You should *never* refer to $\Delta\tau_C$ as the "gravitational proper time" even though you may be tempted. Yes, it is an extreme value (i.e. the longest time measured by any observer), but proper time is the shortest time measured for a single world line. Remember, we're measuring time on three different world lines, so it isn't the same thing. In fact, a careful look shows $\Delta\tau_C = \Delta t$, which means the distant observer actually measures *coordinate* time.

8.6 Geodesics

Knowing how spacetime curves is great, but our real interest lies in how an object or particle will respond to that curvature. In Section 8.3, we even used the principle of stationary action (Eq. 8.3.2) on a particle to derive Einstein's equation (Eq. 8.2.6). We don't see fields or spacetime curvature, so we can't really take direct measurements. It's the behavior of the matter that we really study.

Flat Spacetime

Classically, a particle's behavior is found using either Newton's second law (Eq. 4.2.6) or Lagrange's equation (Eq. 4.2.14) to determine it's equations of motion. We've already generalized Newton's second law for relativity with 4-force (Eq. 7.4.26), which looked a little like this:

$$F^\delta = m_p a^\delta = m_p \frac{du^\delta}{d\tau} = m_p \frac{d^2 x^\delta}{d\tau^2},$$

where a^δ is 4-acceleration and u^δ is 4-velocity. The rest mass, m_p (i.e. the smallest measurable mass), and the proper time, τ (i.e. the shortest measurable time), were first defined at the end of Section 7.2. Usually, if we're trying to find equations of motion, then we write this as

$$\frac{d^2 x^\delta}{d\tau^2} = \frac{F^\delta}{m_p} \tag{8.6.1}$$

so we have just the motions on the left. If the particle or object has no forces acting on it, then we call it a **free particle**. In this case, Newton's second law reduces to

$$\frac{d^2 x^\delta}{d\tau^2} = 0, \tag{8.6.2}$$

which is something akin to Newton's first law. A particle under these conditions would travel in a straight line (i.e. the shortest distance between two points) at constant velocity.

Time-like Geodesics

Until general relativity, gravity was always considered a force, but it didn't quite behave like the others we knew about. Sure, the mathematical descriptions are similar in form as we saw with Coulomb's law (Eq. 5.2.1) and Newton's universal law of gravitation (Eq. 5.2.2). However, when you actually apply these in Newton's second law (Eq. 4.2.6), they behave very differently. The mass and charge are both important when determining the electric influence on an object. When it comes to the gravitational influence though, *neither* is necessary. All that matters (pun intended) for gravity is how the object is moving and it's distance from the source of the gravity. It's weird!!

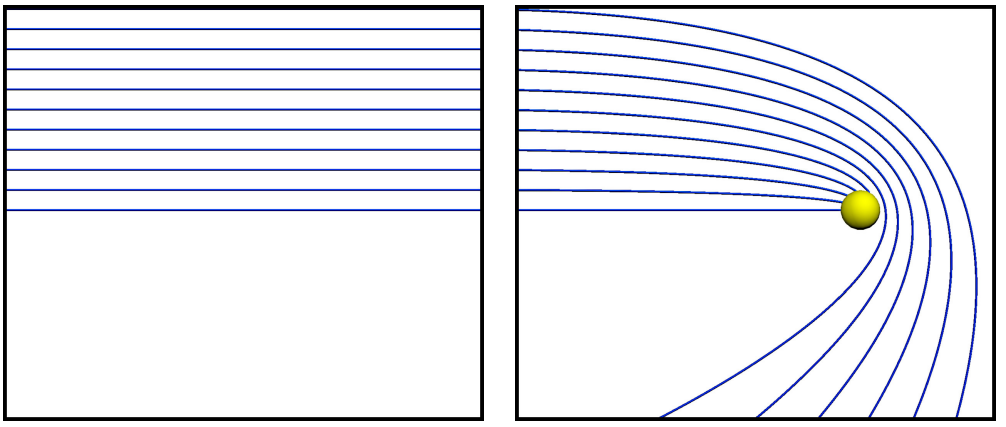


Figure 8.6: These two diagrams feature the same 11 geodesic paths in a particular region of space (just a sample of the infinite number of them). On the left, is a flat spacetime (i.e. the spacetime far away from any sources of curvature). On the right, a massive object like a star is present, so the geodesics are not what we would consider "straight." Also, keep in mind, geodesics are speed dependent, so these curves would be "straighter" for faster moving objects.

But now, gravity is simply the result of curved spacetime, so it's weirdness makes a lot more sense. Since it's no longer considered a force, a particle can be under the influence of gravity and still be considered "free." Unfortunately, by observation, we know these types of particles do *not* travel in what we would think of as "straight" lines as they did in classical physics. This discrepancy can only be resolved if we relax our definition of the word "straight."

To avoid confusion, the new notion of a straight line is called a **geodesic**. In flat spacetime, far away from any massive objects, a geodesic is very straight and obeys Eq. 8.6.2 (see left image in Figure 8.6). However, the lines (or paths) become curved when the spacetime is curved by a massive object like the Earth or the Sun (see right image in Figure 8.6). They might not obey Eq. 8.6.2, but geodesic paths always obey the following definition:

- **Geodesic path** - Any world line between two events such that the proper time is extreme (i.e. maximum or minimum),

where is consistent with the classical definition since the shortest distance takes the least time. Any world line, as defined in Section 7.2, has a proper

time measured in the frame of the particle traveling along the world line. A geodesic path is simply a world line with the best value for proper time.

Powered by this idea of geodesics, we'll need to generalize Eq. 8.6.2 so we can find equations of motion for the particle. The direction of a path is described by the 4-velocity, u^δ, of a particle on that path since that vector is always tangent to the path. For a geodesic path, we can say

$$u^\mu \, \nabla_\mu u^\delta = 0,$$

where $\nabla_\mu u^\delta$ is the change in the δ component of the 4-velocity in the x^μ direction. Multiplying this be u^μ gives us something like a dot product (Eq. 2.2.1), so, essentially, we're saying u^μ is always perpendicular to its change along a geodesic path. In other words, particles traveling on a geodesic path don't change their motion in the direction of their motion.

Using the definition of the covariant derivative on contravariant vectors (Eq. 6.7.3), we get

$$u^\mu \left(\frac{\partial u^\delta}{\partial x^\mu} + \Gamma^\delta_{\mu\nu} u^\nu \right) = 0$$

$$u^\mu \frac{\partial u^\delta}{\partial x^\mu} + \Gamma^\delta_{\mu\nu} u^\mu u^\nu = 0.$$

By the chain rule for derivatives (Eq. 3.1.2), we get

$$u^\mu \frac{du^\delta}{d\tau} \frac{\partial \tau}{\partial x^\mu} + \Gamma^\delta_{\mu\nu} u^\mu u^\nu = 0$$

$$u^\mu \frac{du^\delta}{d\tau} \frac{1}{dx^\mu / d\tau} + \Gamma^\delta_{\mu\nu} u^\mu u^\nu = 0$$

and, since $u^\mu = dx^\mu / d\tau$ (Eq. 7.4.3), this becomes

$$\frac{du^\delta}{d\tau} + \Gamma^\delta_{\mu\nu} u^\mu u^\nu = 0.$$

Note that partial and full derivatives with respect to proper time are equivalent (a quality we've used a lot in this book). This can be written with 4-acceleration in its familiar form using Eq. 7.4.3 again, arriving at

$$\frac{d^2 x^\delta}{d\tau^2} + \Gamma^\delta_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau} = 0 \,, \tag{8.6.3}$$

which is sometimes referred to as the **geodesic equation**. It should be clarified that Eq. 8.6.3 is really only accurate when the particle (or object) being studied does *not* significantly affect the spacetime curvature. Usually this isn't a problem because of the drastic difference in mass we see between people and planets or between planets and stars. However, if we're studying a binary star system, we'd have to be a little more careful.

You can get Eq. 8.6.3 more rigorously by applying a variation principle on proper time,

$$\tau = \int d\tau \quad \Rightarrow 0 = \delta\tau = \delta \int d\tau,$$

and applying the line element (Eq. 7.2.3),

$$0 = \delta \int \sqrt{\frac{d\tau^2}{d\tau d\tau}}\, d\tau = \delta \int \sqrt{\frac{-ds^2}{d\tau d\tau}}\, d\tau = \delta \int \sqrt{-g_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau}}\, d\tau.$$

In the process of arriving at Eq. 8.6.3, we would inadvertently derive our original definition of the Christoffel symbols (Eq. 6.7.6).

Example 8.6.1

Determine the value of k_2 in Eq. 8.5.13 from Example 8.5.5 using the geodesic equation (Eq. 8.6.3).

- We're going to keep things as simple as possible without making any unnecessary approximations. Let's assume the event we're considering (for the geodesic equation) is a *release* event. The small object ($m_{\text{obj}} << m_{\text{star}}$) is being released from rest some distance above the star. As you would expect, this object would experience an acceleration radially toward the star, so we'll consider the $\delta = r$ component:

$$\frac{d^2r}{d\tau^2} + \Gamma^r_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0$$

$$\frac{d^2r}{d\tau^2} = -\Gamma^r_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau}.$$

- Since we're assuming a *release* event, we know

$$\left.\frac{dr}{d\tau}\right|_{\text{event}} = \left.\frac{d\theta}{d\tau}\right|_{\text{event}} = \left.\frac{d\phi}{d\tau}\right|_{\text{event}} = 0$$

and the geodesic equation becomes

$$\frac{d^2r}{d\tau^2} = -\Gamma^r_{tt}\frac{dt}{d\tau}\frac{dt}{d\tau} + 0 + 0 + \dots,$$

where the other 15 terms in the summations are zero. Multiplying through by $dt^2/d\tau^2$, we get

$$\frac{d^2r}{dt^2} = -\Gamma^r_{tt},$$

which represents the acceleration with respect to coordinate time.

- This Christoffel symbol is given in Section C.5 and the form of the metric components is given in Eq. 8.5.13, so

$$\frac{d^2r}{dt^2} = -\frac{1}{2b}\frac{\partial a}{\partial r} = -\frac{1}{2}\left(1 - \frac{k_2}{r}\right)\frac{\partial}{\partial r}\left[-\left(1 - \frac{k_2}{r}\right)\right]$$

$$\frac{d^2r}{dt^2} = -\frac{1}{2}\left(1 - \frac{k_2}{r}\right)\left(-\frac{k_2}{r^2}\right) = \frac{k_2}{2r^2}\left(1 - \frac{k_2}{r}\right)$$

$$\frac{d^2r}{dt^2} = \frac{k_2}{2r^2} - \frac{k_2^2}{2r^3}.$$

- Now we're going to make one approximation in r. This will not affect the value of k_2 because we know it's not a function of r. We already made $r \to \infty$ in Example 8.5.5 to find k_1. However, k_2 contains some information about gravity, so we don't want spacetime *completely* flat, just *close* to flat. We'll assume we're releasing from a point where r is large, but not infinite. Since $1/r^3$ approaches zero faster than $1/r^2$, we can say

$$\frac{d^2r}{dt^2} \approx \frac{k_2}{2r^2}.$$

- At this point, we're in an environment where classical gravity should give the same result. Classical gravity (i.e. **Newtonian gravity**) is

given by Eq. 5.2.2 and, when combined with Newton's second law (Eq. 4.2.6), yields an acceleration of

$$-ma = -G\frac{Mm}{r^2} \quad \Rightarrow a = G\frac{M}{r^2}.$$

Converting to geometrized units (see Table 8.1) and writing the acceleration as a derivative, that's

$$\frac{d^2r}{dt^2} = \frac{M}{r^2}$$

and, by comparison to our large r result, we get

$$\frac{k_2}{2r^2} = \frac{M}{r^2} \quad \Rightarrow \boxed{k_2 = 2M}.$$

We've made no assumption about the fluid nature of the matter of the star in deriving this result.

Example 8.6.2

What are the conserved quantities in the Schwarzchild geometry (Eq. 8.5.12)?

- Conserved quantities are usually the result of some kind of symmetry. We know the Schwarzchild geometry is spherically symmetric (in θ and ϕ) and symmetric in time (t), but not radially (r). However, we also know that Schwarzchild geodesics are always in a single plane, so we'll simplify matters by sticking to the xy-plane (i.e. $\theta = \pi/2$). This leaves us with two possible routes for conserved quantities: the t and ϕ components of the geodesic equation (Eq. 8.6.3).

- We'll start with the t-component, which is

$$\frac{d^2t}{d\tau^2} + \Gamma^t_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0.$$

According to the list in Section C.3, there is only one *unique* Christoffel symbol with t as the upper index, so

$$\frac{d^2t}{d\tau^2} + 2\Gamma^t_{tr}\frac{dt}{d\tau}\frac{dr}{d\tau} = 0,$$

where the 2 appears because $\Gamma^\delta_{\mu\nu} = \Gamma^\delta_{\nu\mu}$. Substituting from Section C.3, we get

$$\frac{d^2t}{d\tau^2} + 2\left[\frac{M}{r^2}\left(1-\frac{2M}{r}\right)^{-1}\right]\frac{dt}{d\tau}\frac{dr}{d\tau} = 0$$

$$\frac{d^2t}{d\tau^2} = -\frac{2M}{r^2}\left(1-\frac{2M}{r}\right)^{-1}\frac{dt}{d\tau}\frac{dr}{d\tau}. \tag{8.6.4}$$

- To keep things looking simple for the math ahead, we'll define $\dot{t} \equiv dt/d\tau$ because we're not going to find t anyway. This simplifies Eq. 8.6.4 to

$$\frac{d\dot{t}}{d\tau} = -\frac{2M}{r^2}\left(1-\frac{2M}{r}\right)^{-1}\dot{t}\,\frac{dr}{d\tau}$$

$$\frac{d\dot{t}}{\dot{t}} = -\frac{2M}{r^2}\left(1-\frac{2M}{r}\right)^{-1}dr.$$

The right side can be simplified further with a convenient change of variable. If we say

$$u = 1-\frac{2M}{r} \quad \Rightarrow du = \frac{2M}{r^2}dr,$$

then

$$\frac{d\dot{t}}{\dot{t}} = -\frac{du}{u} \quad \Rightarrow \int\frac{d\dot{t}}{\dot{t}} = -\int\frac{du}{u}$$

$$\ln\left(\dot{t}\right) = -\ln\left(u\right) + \ln\left(\varepsilon\right),$$

where ε is a unitless constant (i.e. our time-conserved quantity). Moving some things around using logarithm rules, we get

$$\ln\left(\dot{t}\right) = \ln\left(\frac{\varepsilon}{u}\right) \quad \Rightarrow \dot{t} = \frac{\varepsilon}{u}$$

and, substituting back to r,

$$\dot{t} = \frac{dt}{d\tau} = \varepsilon\left(1-\frac{2M}{r}\right)^{-1}. \tag{8.6.5}$$

In flat spacetime (i.e. as $r \to \infty$), the time component of 4-momentum (Eq. 7.4.22) is

$$E_{\text{rel}} = p^t = m_p u^t = m_p \frac{dt}{d\tau},$$

so energy per unit rest mass is

$$\frac{E_{\text{rel}}}{m_p} = \frac{dt}{d\tau},$$

just like Eq. 8.6.5. This is why we called the constant "ε." Solving for the constant, we get

$$\boxed{\varepsilon = \left(1 - \frac{2M}{r}\right)\frac{dt}{d\tau}}, \tag{8.6.6}$$

which is a kind of conservation of energy.

- We have one more though. The ϕ-component of the geodesic equation (Eq. 8.6.3) is

$$\frac{d^2\phi}{d\tau^2} + \Gamma^{\phi}_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0,$$

According to the list in Section C.3, there are two *unique* Christoffel symbol with ϕ as the upper index, so

$$\frac{d^2\phi}{d\tau^2} + 2\Gamma^{\phi}_{r\phi}\frac{dr}{d\tau}\frac{d\phi}{d\tau} + 2\Gamma^{\phi}_{\theta\phi}\frac{d\theta}{d\tau}\frac{d\phi}{d\tau} = 0,$$

where the 2's appear because $\Gamma^{\delta}_{\mu\nu} = \Gamma^{\delta}_{\nu\mu}$. However, we've mentioned already that we're staying in the xy-plane, so $d\theta/d\tau = 0$ and

$$\frac{d^2\phi}{d\tau^2} + 2\Gamma^{\phi}_{r\phi}\frac{dr}{d\tau}\frac{d\phi}{d\tau} = 0.$$

Substituting from Section C.3, we get

$$\frac{d^2\phi}{d\tau^2} + 2\left[\frac{1}{r}\right]\frac{dr}{d\tau}\frac{d\phi}{d\tau} = 0$$

$$\frac{d^2\phi}{d\tau^2} = -\frac{2}{r}\frac{dr}{d\tau}\frac{d\phi}{d\tau}. \tag{8.6.7}$$

- To keep things looking simple for the math ahead, we'll define $\dot{\phi} \equiv d\phi/d\tau$ because we're not going to find ϕ anyway. This simplifies Eq. 8.6.7 to

$$\frac{d\dot{\phi}}{d\tau} = -\frac{2}{r}\frac{dr}{d\tau}\,\dot{\phi}$$

$$\frac{d\dot{\phi}}{\dot{\phi}} = -2\frac{dr}{r} \quad \Rightarrow \int \frac{d\dot{\phi}}{\dot{\phi}} = -2\int \frac{dr}{r}$$

$$\ln\left(\dot{\phi}\right) = -2\ln\left(r\right) + \ln\left(\ell\right),$$

where ℓ is a constant (i.e. our phi-conserved quantity). Moving some things around using logarithm rules, we get

$$\ln\left(\dot{\phi}\right) = \ln\left(\frac{\ell}{r^2}\right)$$

$$\dot{\phi} = \frac{d\phi}{d\tau} = \frac{\ell}{r^2}. \tag{8.6.8}$$

In classical physics, the components of angular momentum are given by Eq. 6.6.6, but the general idea is $L = I\omega = mr^2\omega$ so angular momentum per unit rest mass is

$$\frac{L}{m_p} = r^2\omega = r^2\frac{d\phi}{d\tau}.$$

This is why we called the constant "ℓ." Solving Eq. 8.6.8 for the constant, we get

$$\boxed{\ell = r^2\frac{d\phi}{d\tau}}, \tag{8.6.9}$$

which is a kind of conservation of angular momentum.

Null Geodesics

Geodesic paths taken by massless particles are called **null geodesics**. Yes, I said massless. Gravity is *not* the result of a magical force between masses, although that was a good stepping stone for physics and it worked very well for a long time. No, gravity is the result of straight lines not being straight, so everything capable of motion is affected by gravity. This includes massless particles like photons.

As we saw in Section 7.6, we have issues with using proper time, τ, as a parameter in our equations when it comes to particles that travel on null world lines like photons (and all other massless particles). Particles that travel at *exactly* c have zero proper time, so we needed to choose a different parameter. We chose an **affine parameter**, Ω, which maintained the form of all our equations. Using the same process here, the geodesic (Eq. 8.6.3) becomes

$$\frac{d^2x^\delta}{d\Omega^2} + \Gamma^\delta_{\mu\nu}\frac{dx^\mu}{d\Omega}\frac{dx^\nu}{d\Omega} = 0\,, \tag{8.6.10}$$

where we've just replaced all the τ's with Ω's. The quantity Ω is not unique like proper time, so it's a bit more abstract.

This substitution must be done in all our definitions of 4-vectors. No matter what parameter you choose, remember we must *always* get

$$u_\delta u^\delta = a_\delta a^\delta = F_\delta F^\delta = p_\delta p^\delta = 0$$

because they're all null 4-vectors. It's clear from Eq. 8.6.10 that 4-velocity and 4-acceleration are

$$u^\delta = \frac{dx^\delta}{d\Omega} \text{ and } a^\delta = \frac{d^2x^\delta}{d\Omega^2}.$$

There is also a new form of the 4-force

$$F^\delta = \frac{dp^\delta}{d\Omega},$$

in terms of 4-momentum. With 4-momentum, we have to be a little careful since rest mass, m_p, is zero. It's usually best to define the 4-momentum of the massless particle in terms of the energy rather than worry about its derivative definition like we did with Eq. 7.6.3 (recall that $E_{\text{rel}} = p_{\text{rel}}c$ and $E_{\text{rel}} = hf_{\text{rel}}$ for a photon).

Non Geodesics

If, for some reason, your scenario involves more influences than just gravity, then we need to refer back to Eq. 8.6.1. On the right side of the equation, we had a force term, which was not necessary before because gravity is no longer a force. Adding this back in, Eq. 8.6.3 becomes

$$\frac{d^2x^\delta}{d\tau^2} + \Gamma^\delta_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = \frac{F^\delta}{m_p}, \tag{8.6.11}$$

where F^δ is the total 4-force. We've left the Christoffel term on the left side this whole time because we're still keeping motions on the left and forces on the right (as was done in the flat spacetime case).

A common example of a force affecting an object is the electromagnetic force (assuming it's also charged). Given Eq. 8.6.11, you'd just refer back to the Lorentz 4-force (Eq. 7.5.30), given by

$$F^\delta = qu_\alpha \mathcal{F}^{\delta\alpha},$$

but we need the indices in the correct place. Since $u_\alpha = u^\sigma g_{\alpha\sigma}$ and $\mathcal{F}^{\delta\alpha} = g^{\alpha\lambda}\mathcal{F}^\delta_\lambda$, we get

$$F^\delta = q\left(u^\sigma g_{\alpha\sigma}\right)\left(g^{\alpha\lambda}\mathcal{F}^\delta_\lambda\right) = qu^\sigma\,\delta^\lambda_\sigma\mathcal{F}^\delta_\lambda = qu^\sigma\mathcal{F}^\delta_\sigma,$$

where $\delta^\lambda_\sigma = g_{\alpha\sigma}g^{\alpha\lambda}$ is the Kronecker delta (Eq. 6.2.2). Now, the equation of motion (Eq. 8.6.11) becomes

$$\frac{d^2x^\delta}{d\tau^2} + \Gamma^\delta_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = \frac{qu^\sigma\mathcal{F}^\delta_\sigma}{m_p}$$

and, by the definition of the 4-velocity (Eq. 7.4.3), we get

$$\frac{d^2x^\delta}{d\tau^2} + \Gamma^\delta_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = \frac{q}{m_p}\frac{dx^\sigma}{d\tau}\mathcal{F}^\delta_\sigma, \tag{8.6.12}$$

where q is the charge of the particle and $\mathcal{F}^\delta_\sigma = g_{\alpha\sigma}\mathcal{F}^{\delta\alpha}$ is the mixed EM field tensor (Eq. 7.5.12) affecting the particle.

8.7 Limits and Limitations

When we take general relativity to its limits, things get a little *extreme* (pun intended). What I mean is the theory has places where it fails, but it takes some very extreme circumstances to happen. The big two are black holes and cosmology, so I'll spend this section discussing them and end the chapter with a *bang* (yep, I said it).

Black Holes

In Section 8.4, we mentioned something called a **black hole**, an object so massive that not even light can escape. This seems pretty extreme since the speed of light is the fastest anything can go. They occur when very massive stars run out of material to fuse and collapse. There is so much mass that no force known to us is strong enough to overpower gravity. That includes the forces involved in keeping matter from occupying the same space at the same time.

If a black hole is static (i.e. non-rotating), then the Schwarzchild solution (Eq. 8.5.12) is sufficient in describing it. However, it has issues. A **singularity** is a purely mathematical term describing a place where a function is undefined. Eq. 8.5.12 happens to have two of these, one at $r = 2M$ in g_{rr} and another at $r = 0$ in both g_{tt} and g_{rr}.

The singularity at $r = 2M$ is only a *coordinate* singularity, which means it only exists because of our choice of coordinates. A coordinate transformation we can use to eliminate it is

$$t = t^* - 2M \ln \left| \frac{r}{2M} - 1 \right|, \tag{8.7.1}$$

where t^* is replacing t. Its derivative is

$$dt = dt^* - 2M \left(\frac{r}{2M} - 1 \right)^{-1} \frac{dr}{2M} = dt^* - \left(\frac{r}{2M} - 1 \right)^{-1} dr$$

$$dt = dt^* - \frac{2M}{r} \left(1 - \frac{2M}{r} \right)^{-1} dr$$

However, it's dt^2 in the line element, so

$$dt^2 = (dt^*)^2 - \frac{4M}{r} \left(1 - \frac{2M}{r} \right)^{-1} dt^* dr + \frac{4M^2}{r^2} \left(1 - \frac{2M}{r} \right)^{-2} dr^2.$$

Even better, the entire time component is

$$-\left(1-\tfrac{2M}{r}\right)dt^2 = -\left(1-\tfrac{2M}{r}\right)(dt^*)^2 + \tfrac{4M}{r}dt^*dr - \tfrac{4M^2}{r^2}\left(1-\tfrac{2M}{r}\right)^{-1}dr^2$$

and, with a little algebra, that last term becomes

$$\begin{aligned} -\tfrac{4M^2}{r^2}\left(1-\tfrac{2M}{r}\right)^{-1}dr^2 &= \left[1-\tfrac{4M^2}{r^2}-1\right]\left(1-\tfrac{2M}{r}\right)^{-1}dr^2 \\ &= \left[\left(1-\tfrac{2M}{r}\right)\left(1+\tfrac{2M}{r}\right)-1\right]\left(1-\tfrac{2M}{r}\right)^{-1}dr^2 \\ &= \left(1+\tfrac{2M}{r}\right)dr^2 - \left(1-\tfrac{2M}{r}\right)^{-1}dr^2. \end{aligned}$$

Making this substitution in Eq. 8.5.12 give us

$$ds^2 = -\left(1-\frac{2M}{r}\right)(dt^*)^2 + \frac{4M}{r}dt^*dr + \left(1+\frac{2M}{r}\right)dr^2 + \ldots, \tag{8.7.2}$$

where the angular components remain unchanged. Notice there is no longer a singularity at $r = 2M$ since $1-2M/r$ is no longer in a denominator. This is called **Eddington-Finkelstein coordinates** after Arthur Eddington who invented the transformation in 1924 and David Finkelstein who used it in 1958 to eliminate the coordinate singularity.

This is not to say $r = 2M$ isn't special. It happens to be a place we called the **event horizon**. Imagine you're in a boat on a very calm ocean. Your "horizon" is the farthest you can see or the boundary beyond which you cannot see. The event horizon is a boundary beyond which you cannot see events. All events taking place within the event horizon are incapable of interacting with events outside it. When we discuss the "size" of a black hole, we're referring to the size of the event horizon. In fact, $r = 2M$ is called the **Schwarzchild radius**, which can be written as

$$r_S = \frac{2GM}{c^2} \tag{8.7.3}$$

in SI units rather than geometrized units (converted using Eq. 8.4.3).

Unfortunately, Eq. 8.7.2 still contains the singularity at $r = 0$. This is called a *physical* singularity since it is present with *any* set of coordinates. A single counterexample won't suffice as a proof, so we'll need something a little more encompassing. Traditionally, we find the value of

$$K = R^{\lambda\alpha\mu\nu}R_{\lambda\alpha\mu\nu} \tag{8.7.4}$$

which is a full contraction of the covariant Riemann curvature tensor ($R_{\lambda\alpha\mu\nu} = g_{\lambda\delta}R^{\delta}_{\alpha\mu\nu}$) to a scalar. It is often called the **Kretschmann invariant** because it is a spacetime invariant.

Finding the Kretschmann invariant can be a bit tedious though. There are four summations, which means $4^4 = 256$ terms added together, one for every component of the Riemann tensor. We should be able to reduce this number with a little knowledge about symmetries. Another problem is that we only have the components for the mixed form of Riemann tensor in the Appendix C, not the complete covariant form. If we play with the indices a little, then we get

$$K = R^{\lambda\alpha\mu\nu}R_{\lambda\alpha\mu\nu} = \left(g^{\alpha\rho}g^{\mu\sigma}g^{\nu\eta}R^{\lambda}_{\rho\sigma\eta}\right)\left(g_{\lambda\delta}R^{\delta}_{\alpha\mu\nu}\right)$$

$$K = g^{\alpha\rho}g^{\mu\sigma}g^{\nu\eta}g_{\lambda\delta}R^{\lambda}_{\rho\sigma\eta}R^{\delta}_{\alpha\mu\nu}.$$

At first glance, it may seem as though we've made things worse since this has $4^8 = 65,536$ terms. However, we know $g_{\lambda\delta}$ and $g^{\alpha\rho}$ are both diagonal in the Schwarzchild solution (Eq. 8.5.12). This means $\alpha = \rho$, $\mu = \sigma$, $\nu = \eta$, and $\lambda = \delta$; and we get

$$K = g^{\alpha\alpha}g^{\mu\mu}g^{\nu\nu}g_{\delta\delta}R^{\delta}_{\alpha\mu\nu}R^{\delta}_{\alpha\mu\nu}$$

$$K = g^{\alpha\alpha}g^{\mu\mu}g^{\nu\nu}g_{\delta\delta}\left(R^{\delta}_{\alpha\mu\nu}\right)^2.$$

There are a lot of repeated indices, but only the up/down ones get summed over. We've reduced this back to only four summations or $4^4 = 256$ terms.

Since the Riemann tensor has skew symmetry (Eq. 8.1.2) in the last two indices and any negative signs cancel due to $\left(R^{\delta}_{\alpha\mu\nu}\right)^2$, we can say

$$K = 2g^{\alpha\alpha}g^{\mu\mu}g^{\nu\nu}g_{\delta\delta}\left(R^{\delta}_{\alpha\mu\nu}\right)^2,$$

where the 2 in front accounts for the repeats of $\mu\nu$ and we've reduced to 128 terms. Based on the Riemann components listed in Section C.3, we also know the indices always alternate. That means $\delta \neq \alpha$, $\mu \neq \nu$, $\delta = \mu$, and $\alpha = \nu$; and we get

$$K = 2\sum_{\nu}\sum_{\mu} g^{\nu\nu}g^{\mu\mu}g^{\nu\nu}g_{\mu\mu}\left(R^{\mu}_{\nu\mu\nu}\right)^2,$$

where I've included the summation signs for clarity at this point. There are too many repeated indices! Since $g^{\mu\mu} = 1/g_{\mu\mu}$ for any μ (because they are inverse diagonal tensors), they cancel as well and we're left with

$$K = 2\sum_{\nu}\sum_{\mu}\left(g^{\nu\nu}\right)^2\left(R^{\mu}_{\nu\mu\nu}\right)^2.$$

Note, in this notation, $R^{\mu}_{\nu\mu\nu}$ is *not* the Ricci tensor since the summation occurs *after* it's squared.

Only two summations left makes for $4^2 = 16$ terms, but we know $\mu \neq \nu$ meaning we only have $16 - 4 = 12$ terms left. This is conveniently the exact number of Riemann components given in Section C.3. Expanding the summations gives us

$$K = 2\sum_{\nu}\left(g^{\nu\nu}\right)^2\left[\left(R^{t}_{\nu t\nu}\right)^2 + \left(R^{r}_{\nu r\nu}\right)^2 + \left(R^{\theta}_{\nu\theta\nu}\right)^2 + \left(R^{\phi}_{\nu\phi\nu}\right)^2\right]$$

$$\begin{aligned} K \;=\; & 2\left(g^{tt}\right)^2\left(R^{t}_{rtr}\right)^2 + 2\left(g^{tt}\right)^2\left(R^{t}_{\theta t\theta}\right)^2 + 2\left(g^{tt}\right)^2\left(R^{t}_{\phi t\phi}\right)^2 \\ & +2\left(g^{rr}\right)^2\left(R^{r}_{trt}\right)^2 + 2\left(g^{rr}\right)^2\left(R^{r}_{\theta r\theta}\right)^2 + \left(g^{rr}\right)^2\left(R^{r}_{\phi r\phi}\right)^2 \\ & +2\left(g^{\theta\theta}\right)^2\left(R^{\theta}_{t\theta t}\right)^2 + 2\left(g^{\theta\theta}\right)^2\left(R^{\theta}_{r\theta r}\right)^2 + 2\left(g^{\theta\theta}\right)^2\left(R^{\theta}_{\phi\theta\phi}\right)^2 \\ & +2\left(g^{\phi\phi}\right)^2\left(R^{\phi}_{t\phi t}\right)^2 + 2\left(g^{\phi\phi}\right)^2\left(R^{\phi}_{r\phi r}\right)^2 + 2\left(g^{\phi\phi}\right)^2\left(R^{\phi}_{\theta\phi\theta}\right)^2 \end{aligned}$$

and, substituting in the components of the Riemann tensor and the inverse metric, we get

$$\begin{aligned} K \;=\; & 2\left(4M/r^6\right) + 2\left(M/r^6\right) + 2\left(M/r^6\right) \\ & +2\left(4M/r^6\right) + 2\left(M/r^6\right) + 2\left(M/r^6\right) \\ & +2\left(M/r^6\right) + 2\left(M/r^6\right) + 2\left(4M/r^6\right) \\ & +2\left(M/r^6\right) + 2\left(M/r^6\right) + 2\left(4M/r^6\right) \end{aligned}$$

$$K = \frac{48M^2}{r^6}, \tag{8.7.5}$$

for the Schwarzchild solution.

Notice, it still contains $r = 0$ as a singularity even though it's a spacetime invariant. No matter how you label that singularity in your coordinates, Eq.

8.7.5 is still undefined there. It is a point of infinite curvature where all the mass of a black hole is located and the point where general relativity is insufficient to describe what's happening. Most physicists avoid dealing with this singularity by realizing that anything that happens there is behind an event horizon and, therefore, has no influence over anything that happens in the normal universe. I think that's kind of a cop-out, but it works until someone comes up with a better solution.

Neither of these singularities are an issue for normal stars because of where the matter is located. For example, the Sun's Schwarzchild radius (Eq. 8.7.3) is $2M = 2(1477 \text{ m}) = 2954$ meters, but its actual radius is 6.955×10^8 meters. Since $r = 2M$ is inside the Sun, we'd use a different line element involving $m(r)$ rather than M. Recall, from Eq. 8.5.10, that $m(r)$ is the mass enclosed by a sphere of radius r (centered at the center of the star). We also know $m(r) \to 0$ as $r \to 0$, which resolves the singularity at $r = 0$. These singularities are only an issue when all the mass of the star is *inside* $r = 2M$.

Example 8.7.1

Describe the path of a photon traveling only radially close to a black hole.

- If an object is only traveling along radial lines, then the Schwarzchild line element (Eq. 8.5.12) simplifies to

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2.$$

  Since we're dealing with photons which travel on null paths, we get say

$$0 = -\left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2.$$

- Now we just have to solve this for t in terms of r. Moving some things around, we get

$$\left(1 - \frac{2M}{r}\right) dt^2 = \left(1 - \frac{2M}{r}\right)^{-1} dr^2 \quad \Rightarrow dt^2 = \left(1 - \frac{2M}{r}\right)^{-2} dr^2$$

$$\Rightarrow dt = \pm \left(1 - \frac{2M}{r}\right)^{-1} dr,$$

  where the square root shows we have two solutions.

- Integrating, this becomes

$$\int dt = \int \pm \left(1 - \frac{2M}{r}\right)^{-1} dr = \int \frac{r}{2M} \left(\frac{r}{2M} - 1\right)^{-1} dr$$

and, substituting $u = r/2M - 1$ and $du = dr/2M$,

$$\int dt = \int \pm \left(\frac{u+1}{u}\right) (2M\, du) = \pm 2M \int \left(1 + \frac{1}{u}\right) du$$

$$t = \pm \left(2M\, u + 2M \ln |u|\right) + \text{constant}.$$

Substituting back into r, we get

$$\boxed{t = \pm \left(r + 2M \ln \left|\frac{r}{2M} - 1\right|\right) + \text{constant}}, \tag{8.7.6}$$

where that big term was used in Eq. 8.7.1 was used to eliminate the coordinate singularity at $r = 2M$. The constant on the end just means the path it describes can take place at any time in the future or the past. For visual purposes, it's usually best to give this constant a non-zero value so the horizontal axis (i.e. the space axis) doesn't pass through any events of interest.

- Plotting our two solutions from Eq. 8.7.6 in a spacetime diagram results in Figure 8.7. The solutions are labeled "plus" and "minus" for the $\pm$. The plus curves shows a photon traveling inward when inside the event horizon (r decreases as t increases) and outward when outside (r increases as t increases), which makes perfect sense. The minus curve is a bit strange though. It describes an inward traveling photon from outside the event horizon and it would appear it never reaches it. Luckily, this is only the result of our choice of coordinates. The Schwarzchild solution has a coordinate singularity there.

- We can transform to Eddington-Finkelstein coordinates using Eq. 8.7.1 to eliminate the coordinate singularity and get a better idea of what's happening. The easiest way to do this is to solve for t^* in Eq. 8.7.1 resulting in

$$t^* = t + 2M \ln \left|\frac{r}{2M} - 1\right|,$$

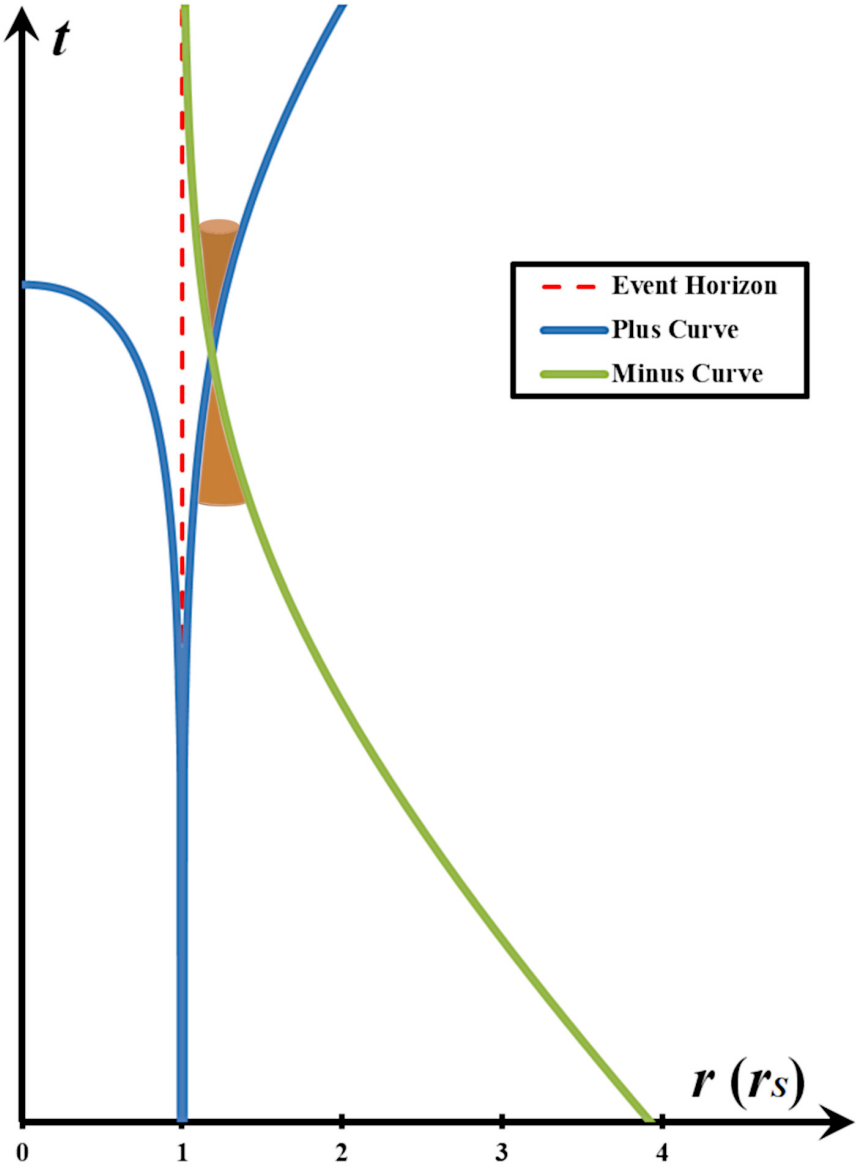


Figure 8.7: This is a spacetime diagram showing two possible worldlines for radially traveling photons. Each of the curves is a solution from Eq. 8.7.6. Since the units of the radial axis are in Schwarzchild radii (i.e. $r_S = 2M$), the event horizon is located at $r = 1\, r_S$. The physical singularity is still located at $r = 0$.

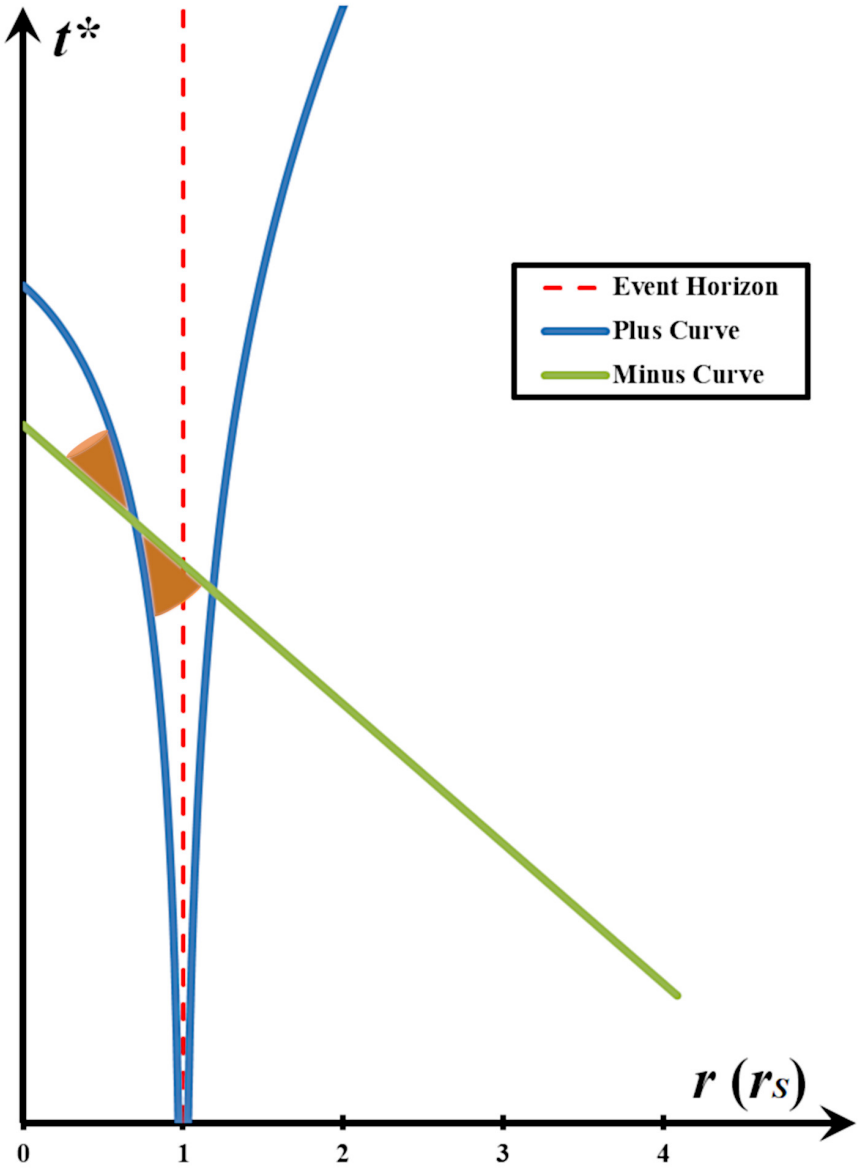


Figure 8.8: This is Figure 8.7 transformed into Eddington-Finkelstein coordinates. It fixes the coordinate singularity at $r = 2M$ allows us to predict how photons (and anything else) will cross the event horizon. A light cone has been shown inside the event horizon for dramatic effect.

so essentially we just need to add that big term to our solutions from Eq. 8.7.6. This gives us

$$t^* = \left[\pm\left(r + 2M\ln\left|\frac{r}{2M} - 1\right|\right) + \text{constant}\right] + 2M\ln\left|\frac{r}{2M} - 1\right|$$

$$\boxed{\left\{\begin{array}{lcl} t_1^* &=& r + 4M\ln\left|\frac{r}{2M} - 1\right| + \text{constant} \\ t_2^* &=& -r + \text{constant}. \end{array}\right\}}$$

This doesn't change the shape of the plus curve very much as you can see in Figure 8.8. However, the minus curve changes dramatically because it is now a straight line. This minus curve clearly crosses the event horizon and makes it all the way to the physical singularity at $r = 0$. These two curves form light cones (see Section 7.2) that progressively point toward the black hole. Even weirder, inside the event horizon time-like world lines become space-like. No one has any real concept of what that even means. It's crazy!

Example 8.7.2

A photon can escape a black hole from just above the event horizon if it travels straight away from it. If there is any angle to its trajectory, then it will fall back into the black hole. The boundary beyond which a photon can escape at *any* angle away from the black hole is the radius at which a photon can *orbit* in a circle. What is this radius?

- There's a lot going on here, so let's get all are ducks in a row. A convenient choice for the plane of the circle is the xy-plane. In spherical coordinates, the xy-plane is defined by $\theta = \pi/2$, from which it follows that

$$\frac{d^2\theta}{dt^2} = \frac{d\theta}{dt} = 0.$$

Furthermore, if the photon's path is a circular orbit, then

$$\frac{dr}{dt} = 0 \text{ and } \frac{d\phi}{dt} = \text{constant}$$

as well as

$$\frac{d^2r}{dt^2} = \frac{d^2\phi}{dt^2} = 0,$$

for all events on the path. Note that we are allowed to take derivatives with respect to *coordinate* time for photon, just not with respect to *proper* time.

- In this case, the Schwarzchild line element (Eq. 8.5.12) simplifies to

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + r^2\, d\phi^2.$$

Since we're dealing with photons which travel on null paths, we get say

$$0 = -\left(1 - \frac{2M}{r}\right) dt^2 + r^2\, d\phi^2.$$

Moving some stuff around, we get

$$r^2\, d\phi^2 = \left(1 - \frac{2M}{r}\right) dt^2$$

$$\left(\frac{d\phi}{dt}\right)^2 = \frac{1}{r^2}\left(1 - \frac{2M}{r}\right). \tag{8.7.7}$$

- We're gong to need another equation to eliminate ϕ, so we'll use the geodesic equation for massless particles (Eq. 8.6.10). We'll only need the radial component, which is

$$\frac{d^2r}{d\Omega^2} + \Gamma^r_{\mu\nu}\frac{dx^\mu}{d\Omega}\frac{dx^\nu}{d\Omega} = 0,$$

where all derivative are with respect to an affine parameter Ω. If we multiply through by $d\Omega^2/dt^2$, then we get

$$\frac{d^2r}{dt^2} + \Gamma^r_{\mu\nu}\frac{dx^\mu}{dt}\frac{dx^\nu}{dt} = 0,$$

where all derivatives are now with respect to coordinate time. Expanding the sum gives a total of $1+4\times4=17$ terms. However, all the zero derivatives mentioned earlier brings this down to

$$\Gamma^r_{tt}\frac{dt}{dt}\frac{dt}{dt}+\Gamma^r_{\phi\phi}\frac{d\phi}{dt}\frac{d\phi}{dt}=\Gamma^r_{tt}+\Gamma^r_{\phi\phi}\left(\frac{d\phi}{dt}\right)^2=0$$

only two non-zero terms. The Christoffel symbols can be found in Section C.3, giving us

$$\frac{M}{r^2}\left(1-\frac{2M}{r}\right)-r\left(1-\frac{2M}{r}\right)\sin^2\theta\left(\frac{d\phi}{dt}\right)^2=0.$$

Using $\sin^2\theta=\sin^2(\pi/2)=1$ and a little algebra shows

$$\left(\frac{d\phi}{dt}\right)^2=\frac{M}{r^3}. \tag{8.7.8}$$

- Combining Eqs. 8.7.7 and 8.7.8, we get

$$\frac{1}{r^2}\left(1-\frac{2M}{r}\right)=\frac{M}{r^3}\quad\Rightarrow 1-\frac{2M}{r}=\frac{M}{r}$$

or $r=3M$. This is one and half times the Schwarzchild radius or

$$\boxed{r_{\text{orbit}}=\frac{3GM}{c^2}} \tag{8.7.9}$$

in SI units rather than geometrized units (converted using Eq. 8.4.3). This is sometimes called a **photon sphere**. No known object or particle has a stable orbit inside the sphere. A photon path on this sphere is shown in Figure 8.9.

Examples 8.7.1 and 8.7.2 show some very convenient special cases, but you might be wondering what the general case looks like. What if you want an angled path for a photon or the path of a massive particle? In that case, you're going to have to solve several components of the geodesic equation (Eq.

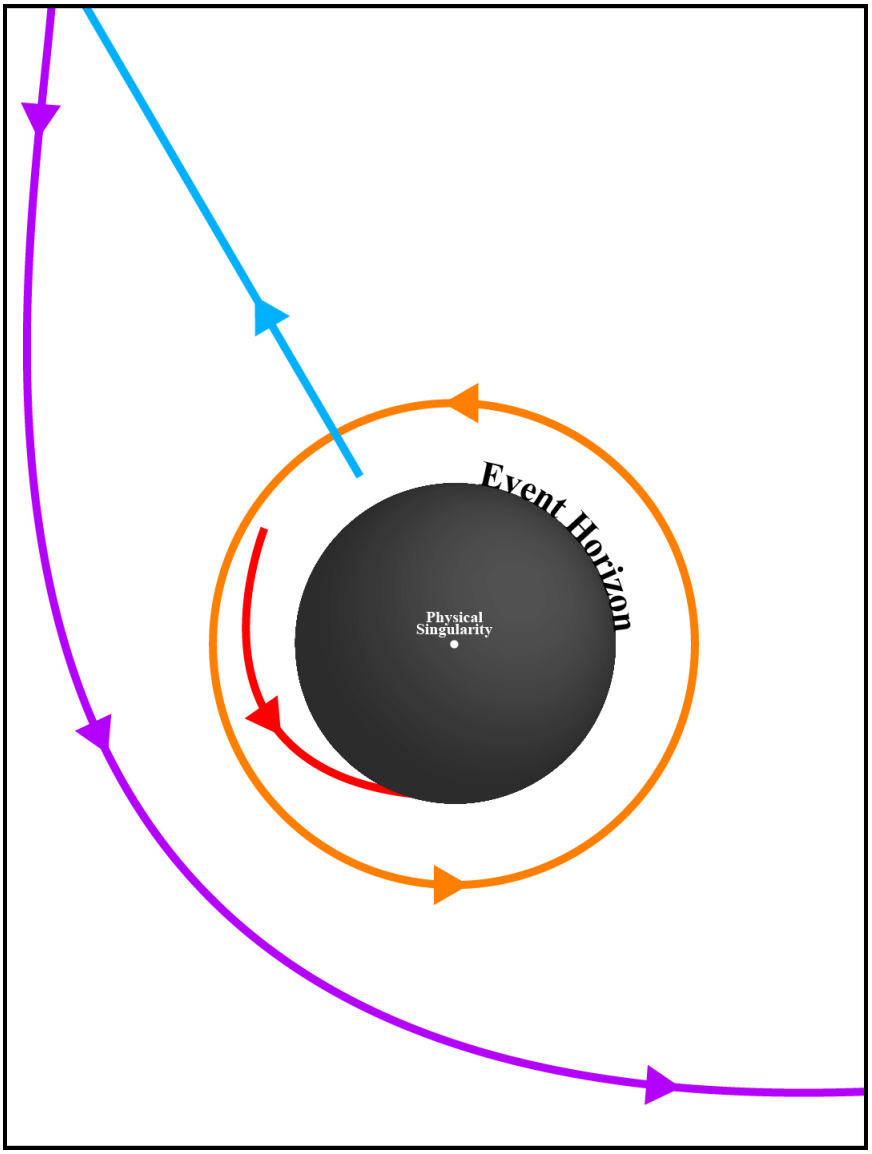


Figure 8.9: These are four different null geodesic paths (i.e. geodesic paths for massless particles like photons) near a black hole. They were plotted using Eqs. 8.7.10 and 8.7.11 with different sets of initial conditions. One of the paths is on the photon sphere described by Eq. 8.7.9 ($r_i = 3M$ and $\ell/\varepsilon = \pm 3\sqrt{3}M$).

8.6.10), which usually requires numerical integration like the Runge-Kutta method from Section A.1.

You should still be able to assume the xy-plane since any path would form in a plane, so you'll only need the r and ϕ components. Furthermore, even for massless particles, Eqs. 8.6.5 and 8.6.8 should still apply as long as you're not using proper time as your independent variable. Assuming again that $\theta = \pi/2$ and its derivatives are zero, we get a general r-component of

$$0 = \frac{d^2r}{d\Omega^2} + \Gamma^r_{tt}\frac{dt}{d\Omega}\frac{dt}{d\Omega} + \Gamma^r_{rr}\frac{dr}{d\Omega}\frac{dr}{d\Omega} + \Gamma^r_{\phi\phi}\frac{d\phi}{d\Omega}\frac{d\phi}{d\Omega}$$

$$\frac{d^2r}{d\Omega^2} = -\Gamma^r_{tt}\frac{dt}{d\Omega}\frac{dt}{d\Omega} - \Gamma^r_{rr}\frac{dr}{d\Omega}\frac{dr}{d\Omega} - \Gamma^r_{\phi\phi}\frac{d\phi}{d\Omega}\frac{d\phi}{d\Omega}$$

Substituting in the Christoffel symbols from Section C.3 as well as Eqs. 8.6.5 and 8.6.8, this becomes

$$\begin{aligned}\frac{d^2r}{d\Omega^2} &= -\frac{M}{r^2}\left(1-\frac{2M}{r}\right)\left[\varepsilon\left(1-\frac{2M}{r}\right)^{-1}\right]^2 \\ &\quad + \frac{M}{r^2}\left(1-\frac{2M}{r}\right)^{-1}\left(\frac{dr}{d\Omega}\right)^2 + r\left(1-\frac{2M}{r}\right)\left(\frac{\ell}{r^2}\right)^2\end{aligned}$$

$$\frac{d^2r}{d\Omega^2} = -\frac{M}{r^2}\left(1-\frac{2M}{r}\right)^{-1}\varepsilon^2 + \frac{M}{r^2}\left(1-\frac{2M}{r}\right)^{-1}\left(\frac{dr}{d\Omega}\right)^2 + \left(1-\frac{2M}{r}\right)\frac{\ell^2}{r^3}$$

$$\boxed{\frac{d^2r}{d\Omega^2} = \frac{M}{r^2}\left(1-\frac{2M}{r}\right)^{-1}\left[\left(\frac{dr}{d\Omega}\right)^2 - \varepsilon^2\right] + \left(1-\frac{2M}{r}\right)\frac{\ell^2}{r^3}}, \tag{8.7.10}$$

where ε and ℓ are constants related to energy and angular momentum, respectively (see Example 8.6.2 for more details). Generalizing Eq. 8.6.8 with Ω gives us

$$\boxed{\frac{d\phi}{d\Omega} = \frac{\ell}{r^2}}, \tag{8.7.11}$$

where, again, ℓ is a constant related to angular momentum. Eqs. 8.7.10 and 8.7.11 apply to both massive ($\Omega = \tau$) and massless particles.

Numerical integration requires you to eliminate all second derivatives (or higher) by increasing the number of variables. In this case, we only have a second derivative in r, so we go from two variables to three:

$$\left\{ \begin{array}{rcl} \dfrac{dr}{d\Omega} & = & u_r \\ \dfrac{du_r}{d\Omega} & = & \dfrac{M}{r^2}\left(1-\dfrac{2M}{r}\right)^{-1}\left[u_r^2-\varepsilon^2\right]+\left(1-\dfrac{2M}{r}\right)\dfrac{\ell^2}{r^3} \\ \dfrac{d\phi}{d\Omega} & = & \dfrac{\ell}{r^2} \end{array} \right\} \tag{8.7.12}$$

giving us all first derivatives on the left and just variables on the right. Initial conditions for u_r are found most easily using $u^\delta u_\delta = g_{\delta\nu}u^\delta u^\nu$, which equals -1 for massive particles (Eq. 7.4.6) and zero for massless particles since they're on null geodesics. For photons, this simplifies to

$$u_r = \pm\sqrt{\varepsilon^2 - \frac{\ell^2}{r^2}\left(1-\frac{2M}{r}\right)}, \tag{8.7.13}$$

which I guarantee will save you a headache. Applying the Runge-Kutta method from Section A.1 for several different initial conditions results in Figure 8.9.

Cosmology and Beyond

Oddly, the general theory of relativity is not just limited to the spacetime around planets and stars. A star is just a huge collection of tiny particles. A galaxy is just a collection of stars and other objects. On an intergalactic level, we can think of a galaxy as a single massive object affecting a much larger spacetime around it. The Milky Way itself has several satellite galaxies. We can further group galaxies into clusters and clusters into superclusters until we reach our ultimate limit: the entire universe.

It was not long after the publishing of general relativity in 1915 that we began extending it like this. Four people (Alexander Friedmann, Georges Lemaître, Howard Robertson, and Arthur Walker), between 1922 and 1935, independently developed a line element for the entire universe. It's called the Friedmann-Lemaître-Robertson-Walker geometry in their honor and is

considered the standard model of cosmology by the scientific community. It takes the form

$$ds^2 = -dt^2 + [a(t)]^2 \left[\frac{1}{1-kr^2}\, dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2 \right], \qquad (8.7.14)$$

where $a(t)$ is the **scale factor** and the constant k is the overall spatial curvature of the universe. The scale factor is defined to be $a = 1$ at the present time and represents the expansion of the universe. We can see it is only on the spatial components, so it is only space that expands, not time. We can think of it as average distance between galactic superclusters:

$$a(t) = \frac{\text{Average Supercluster Spacing}}{400,000,000 \text{ ly}}, \qquad (8.7.15)$$

so the scale factor is unitless because it's normalized to the current spacing of 400 million light years. Since the average supercluster spacing changes with time, so does the scale factor.

The spatial curvature, k, is a different story. It is constant over space *and* time, but its sign has implications:

- **"Closed" Universe** ($k > 0$) - There is (probably) a finite amount of space that curves back on itself. If you travel in straight line for long enough, you'll end up back where you started (like walking around the Earth).

- **"Flat" Universe** ($k = 0$) - There is an infinite amount of space that doesn't curve at all. No matter how long you travel in a straight line, you'll just see more universe.

- **"Open" Universe** ($k < 0$) - There is an infinite amount of space that curves away from itself. If two spaceships start traveling in straight parallel lines, they'll always just see more universe, but they'll drift apart over time (like on the surface of a saddle).

It is often stated that k can only have three values: $+1$, 0, and -1; but this depends on your choice of units. The way Eq. 8.7.14 is written, the quantity kr^2 must be unitless, so k must have units of m^{-2} like a Gaussian curvature. It is not restricted to $+1$, 0, or -1.

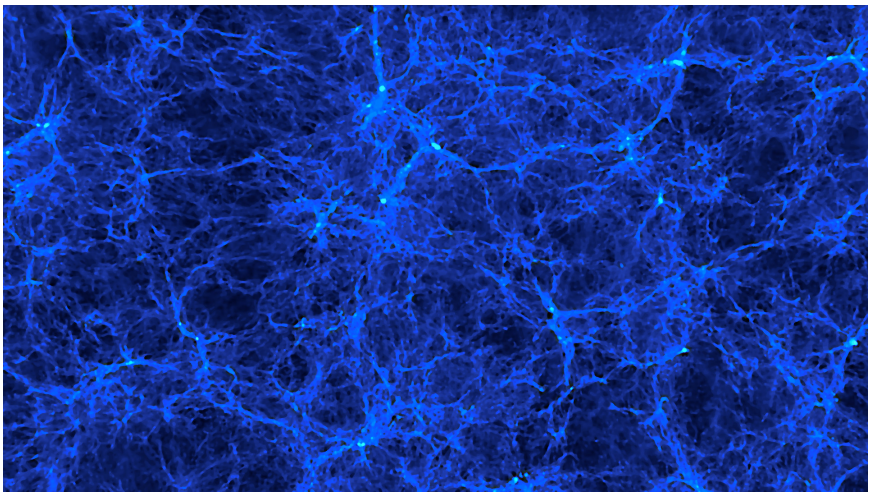


Figure 8.10: This is what the universe looks like on the largest scale. Strings of galactic superclusters stretch across space in a cosmic web leaving large voids between them. (Image credit: Argonne National Laboratory)

If we plan on using this geometry in Einstein's equation (Eq. 8.4.4), then it's more convenient to write it as a metric tensor instead:

$$g_{\alpha\nu} \longrightarrow \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & \frac{a^2}{1-kr^2} & 0 & 0 \\ 0 & 0 & a^2r^2 & 0 \\ 0 & 0 & 0 & a^2r^2\sin^2\theta \end{bmatrix}. \tag{8.7.16}$$

The Ricci curvature scalar and the non-zero components of the Ricci curvature tensor for this geometry can be found in Section C.6 of the appendix. We can see this geometry is spherically symmetric because it matches Eq. 8.5.1 with the exception of the extra factor $a(t)$ (which is fine because a is only a function of time, not space). In fact, this geometry goes further assuming a *perfectly uniform* universe. For Eq. 8.7.14 to apply to a universe *perfectly*, that universe must be both **homogeneous** (the same in every place) and **isotropic** (the same in every direction). While this isn't exactly true for our universe, we can see from Figure 8.10 the universe is uniform *enough* on the largest scale that Eq. 8.7.14 is a very good approximation.

Unfortunately, we have a problem. Since Edwin Hubble in 1929, we've been taking measurements of the expanding universe with increasing accuracy. Data shows that we live in a flat universe ($k = 0$), but there isn't

enough matter and energy in the universe to flatten it. The consequence is we need to adjust Einstein's equation (Eq. 8.4.4) to apply to our universe and we don't have a lot of wiggle room for that. The methods used in Sections 8.2 and 8.3 involve derivatives and integrals, so our only real option in the math is to add a constant, $g_{\alpha\nu}\Lambda$. The generalized Einstein's equation looks like this:

$$R_{\alpha\nu} - \frac{1}{2}g_{\alpha\nu}R + g_{\alpha\nu}\Lambda = 8\pi\, T_{\alpha\nu}\,, \tag{8.7.17}$$

where Λ is called the **cosmological constant**. This constant has units of m^{-2} like an energy density (see Table 8.1), but its *physical* source is unknown, so we call it **dark energy** ("dark" because we're in the dark about it).

Now that we have Eqs. 8.7.16 and 8.7.17, we just need to know what the matter and energy in the universe looks like. On the largest scale of the universe, the matter is uniform and doesn't change much, so let's assume it's a perfect static fluid. Using Eq. 8.5.8 and lowering the indices, we get

$$\left\{\begin{array}{lcl} T_{tt} &=& \rho \\ T_{rr} &=& \left(\frac{a^2}{1-kr^2}\right) P \\ T_{\theta\theta} &=& a^2 r^2 P \\ T_{\phi\phi} &=& a^2 r^2 \sin^2\theta\, P \end{array}\right\} \tag{8.7.18}$$

for the non-zero components of the stress-energy tensor. The quantity ρ is the energy density of the universe, which is a function of time, but not space (because we're assuming the universe is homogeneous). It includes all regular matter, dark matter, and photon energy (basically, anything that isn't dark energy). The quantity P is the pressure inside that energy, which (like ρ) is a function of time, but not space. Essentially, this takes care of any interactions that are happening between parts of that energy.

The ultimate goal of this work is to get a set of equations that describe the past, present, and future of the universe. Einstein's equation (Eq. 8.7.17) results in four independent equations that will give us just that, but we'll do some algebra to bring that down to two. The $\alpha\nu = tt$ component of Einstein's equation gives us

$$R_{tt} - \frac{1}{2}g_{tt}R + g_{tt}\Lambda = 8\pi\, T_{tt}$$

$$\left[-3\frac{\ddot{a}}{a}\right] - \frac{1}{2}\left[-1\right]\left[\frac{6}{a^2}\left(k + \dot{a}^2 + a\ddot{a}\right)\right] + \left[-1\right]\Lambda = 8\pi\left[\rho\right],$$

where dots represent partial derivatives with respect to time. Simplifying, we get

$$-3\frac{\ddot{a}}{a} + 3\frac{k}{a^2} + 3\frac{\dot{a}^2}{a^2} + 3\frac{\ddot{a}}{a} - \Lambda = 8\pi\,\rho$$

$$3\frac{k}{a^2} + 3\frac{\dot{a}^2}{a^2} - \Lambda = 8\pi\,\rho, \tag{8.7.19}$$

which we'll save for later. The $\alpha\nu = rr$, $\theta\theta$, and $\phi\phi$ components of Einstein's equation are all the same because the curvature tensor components, $R_{\alpha\nu}$, are so similar (see Section C.6). The $\theta\theta$ component is the simplest and works out as

$$R_{\theta\theta} - \frac{1}{2}g_{\theta\theta}R + g_{\theta\theta}\Lambda = 8\pi\,T_{\theta\theta}$$

$$\left[r^2\left(2k + 2\dot{a}^2 + a\ddot{a}\right)\right] - \frac{1}{2}\left[a^2r^2\right]\left[\frac{6}{a^2}\left(k + \dot{a}^2 + a\ddot{a}\right)\right] + \left[a^2r^2\right]\Lambda = 8\pi\left[a^2r^2P\right].$$

If we multiply through by $1/(a^2r^2)$, we get

$$\left[\frac{1}{a^2}\left(2k + 2\dot{a}^2 + a\ddot{a}\right)\right] - \frac{1}{2}\left[\frac{6}{a^2}\left(k + \dot{a}^2 + a\ddot{a}\right)\right] + \Lambda = 8\pi\,P$$

and, finally simplifying, gives us

$$2\frac{k}{a^2} + 2\frac{\dot{a}^2}{a^2} + \frac{\ddot{a}}{a} - 3\frac{k}{a^2} - 3\frac{\dot{a}^2}{a^2} - 3\frac{\ddot{a}}{a} + \Lambda = 8\pi\,P$$

$$-\frac{k}{a^2} - \frac{\dot{a}^2}{a^2} - 2\frac{\ddot{a}}{a} + \Lambda = 8\pi\,P. \tag{8.7.20}$$

Unfortunately, having so many factors in Eq. 8.7.20 isn't very convenient. We can simplify further by adding Eq. 8.7.19 to three of Eq. 8.7.20 (i.e. adding all the components together: $tt + rr + \theta\theta + \phi\phi$):

$$\left[3\frac{k}{a^2} + 3\frac{\dot{a}^2}{a^2} - \Lambda\right] + 3\left[-\frac{k}{a^2} - \frac{\dot{a}^2}{a^2} - 2\frac{\ddot{a}}{a} + \Lambda\right] = \left[8\pi\,\rho\right] + 3\left[8\pi\,P\right]$$

$$3\frac{k}{a^2}+3\frac{\dot{a}^2}{a^2}-\Lambda-3\frac{k}{a^2}-3\frac{\dot{a}^2}{a^2}-6\frac{\ddot{a}}{a}+3\Lambda=8\pi\left(\rho+3P\right)$$

$$-6\frac{\ddot{a}}{a}+2\Lambda=8\pi\left(\rho+3P\right). \tag{8.7.21}$$

If we move some things around in Eqs. 8.7.19 and 8.7.21, we get something we can actually interpret. Eq. 8.7.19 determines spatial curvature, k, so solving for that term gives us

$$\frac{k}{a^2}=\frac{8\pi}{3}\rho+\frac{\Lambda}{3}-\frac{\dot{a}^2}{a^2}. \tag{8.7.22}$$

We can see everything inside the universe (regular matter, photons, dark matter, and even dark energy) makes the curvature more positive. The $\dot{a}^2/a^2$ term can be thought of as a kinetic energy (density) term for the universe, which makes the curvature more negative. As mentioned before, our universe appears to be "flat" ($k=0$), so the entire left side equals zero. That leaves us with

$$\frac{\dot{a}^2}{a^2}=\frac{8\pi}{3}\rho+\frac{\Lambda}{3} \tag{8.7.23}$$

for our universe. Eq. 8.7.21 determines the acceleration rate of the universe, $\ddot{a}$, so solving for that term gives us

$$\frac{\ddot{a}}{a}=-\frac{4\pi}{3}\left(\rho+3P\right)+\frac{\Lambda}{3}. \tag{8.7.24}$$

We can see regular matter, photons, and dark matter (ρ and P) all lower the acceleration rate. Dark energy (Λ), on the other hand, raises that acceleration rate.

Eqs. 8.7.22 and 8.7.24 together are called the **Friedmann equations** (after Alexander Friedmann). Solutions to these differential equations, like those found in Figure 8.11, are the scale factor, $a(t)$, which show the expansion of the universe over time. As with any differential equation, those solutions depend either on initial or boundary conditions. The way we defined the scale factor in Eq. 8.7.15, we know $a(\text{now})=1$, so we would just need to know the current values of ρ (energy density) and P (pressure). The

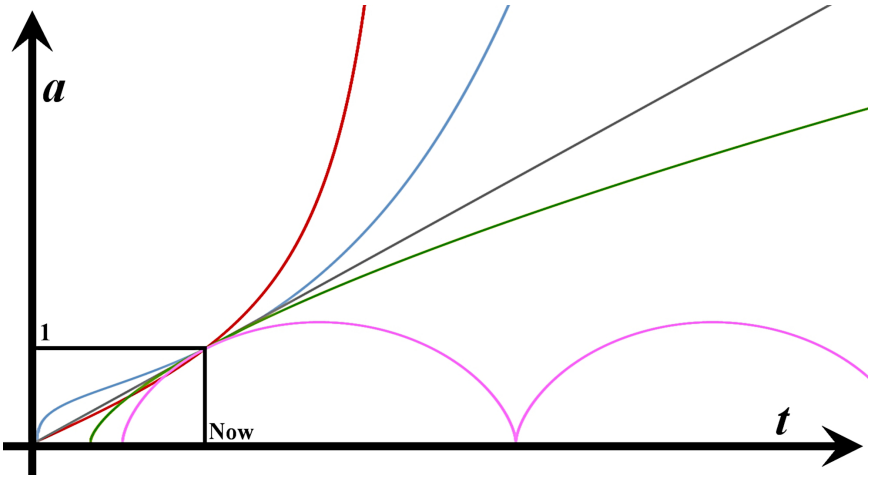


Figure 8.11: This graph shows various possible solutions to the Friedmann equations. The pink universe expands for the first half of its life and contracts again in the second half, ending in a big crunch (or possibly a big bounce). The green universe expands and cools forever, but the rate of that expansion slows over time ending in a big freeze (i.e. a heat death). The gray universe expands forever at a constant rate, which is only possible if the universe is perfectly empty. The blue universe expands forever, slowing for some time then accelerating for the rest and ending in a big freeze (our own universe). The red universe reaches an infinite scale factor, a, in a finite amount of time causing a big rip where the actual fabric of the universe rips to bits.

value of k is constant over time and space and so is the value of Λ (the cosmological constant), which has some weird consequences. For one, ρ and P decrease over time, so the Λ term is eventually the big term even though it doesn't change. That means, as long as $\Lambda > 0$, the last stage of the universe is guaranteed to be an accelerated expansion.

However, we do have a problem when we run the clock backwards. In reverse, the universe gets smaller and smaller until, eventually, the entire (observable) universe becomes a physical singularity (the size of the singularity at the center of a black hole). Recall, at the black hole's singularity, the laws of physics break down because the spacetime curvature is infinite. A similar problem arises here with the universe, but we have no event horizon to hide it behind. General relativity accurately explains the universe every place and every time except the center of black hole and the beginning of the universe, so it would still appear just a bit incomplete. We have yet to find a solution to the problem.

Chapter 9

Basic Quantum Mechanics

9.1 Descent into Madness

Our desire to understand the nature of matter has probably been around as long as we have. In Ancient Greece between 400 and 300 BCE, two camps formed:

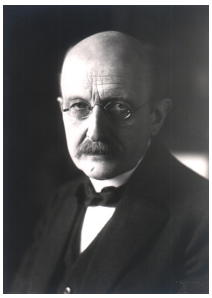
- The followers of Democritus, who believed matter was made of very small pieces they called **atomos** meaning "indivisible," and
- The followers of Aristotle, who believed matter was perfectly continuous.

Aristotle tended to believe the whole universe should fit his vision of perfection and, as a result, he was almost *always* wrong. If you haven't noticed, we get the word **atom** from Democritus and his atomos, so it's clear his camp won this fight. However, it wasn't until the development of electrodynamics (See Chapter 5) in the middle-to-late 19^{th} century, that our technology had advanced to a point where we could start investigating atomic scales. In 1897 CE, a British physicist named J.J. Thomson performed several experiments resulting in the discovery of the electron, kick-starting our descent into madness.

Beginning of Modern Physics

There was a seemingly unrelated problem that had developed in the study of **black body radiation** (i.e. light emitted by objects that do *not* reflect


J.J. Thomson


Max Planck


Ernest Rutherford

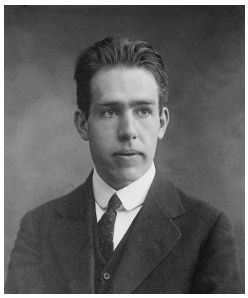

Niels Bohr

Figure 9.1: These people were important in reaching the limits of classical physics.

light, like the Sun). The problem didn't have a name then, but we now call it the **ultraviolet catastrophe** because all the models blew up to infinity in the ultraviolet wavelength range (see Figure 9.2). However, in 1900, a German physicist named Max Planck solved this problem with

$$R_\lambda(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k_B T}} - 1}, \tag{9.1.1}$$

where λ is the wavelength of emitted light, T is the temperature of the object, $h = 6.626 \times 10^{-34}$ Js (or 4.136×10^{-15} eVs) is **Planck's constant**, $k_B = 1.381 \times 10^{-23}$ J/K (or 8.617×10^{-5} eV/K) is Boltzmann's constant, and R_λ is the spectral radiance (i.e. intensity per steradian per unit wavelength). In the process, Planck had to embrace two ideas that made him extremely uncomfortable:

1. The second law of thermodynamics was not fundamental to the universe, but just the result of statistics.

2. The object was composed of very small oscillators.

The result is that light being emitted by the object is not done continuously, but in small packets with specific frequency (or wavelength) called **quanta**. Each "quantum" of energy ($E = hf$) is what we now call a photon of light.

By 1904, J.J. Thomson returned to propose a model for the atom with negative electrons floating in a mist of positive charge (see Figure 9.3). Only one year later, Albert Einstein was using the existence of atoms in several of his famous papers. In 1911, Ernest Rutherford fired alpha radiation into a thin sheet of gold foil. Thomson's "plum pudding" model didn't explain the

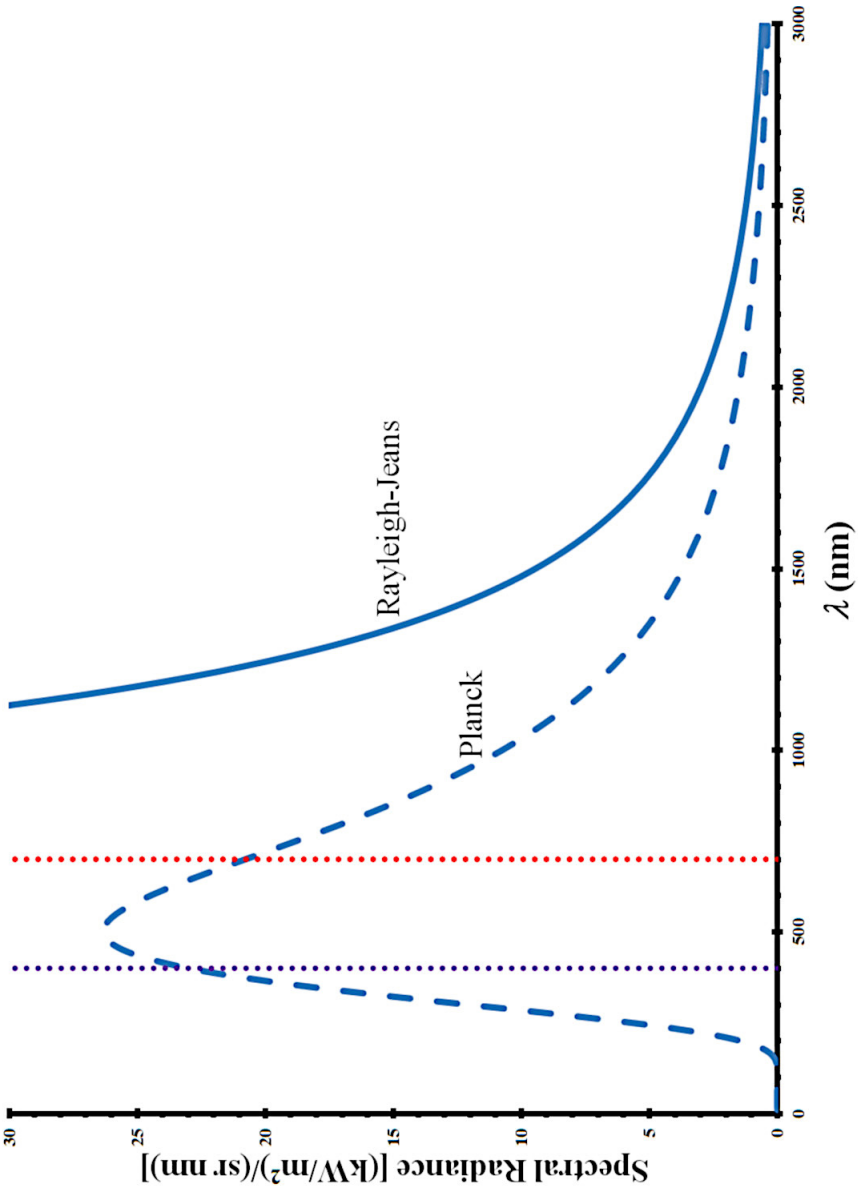


Figure 9.2: This graph shows predicted spectral radiance against wavelength for the sun's surface ($T = 5778$ K). The solid curve is the attempt by John Rayleigh and James Jeans, but fails at low wavelengths like many other attempts. The dashed curve is Planck's solution to the problem.

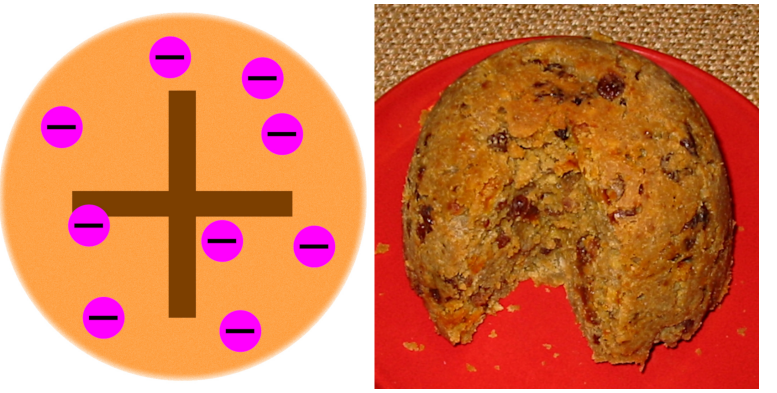


Figure 9.3: On the left is J.J. Thomson's model of the atom. On the right is a picture of plum pudding. During an interview with Thomson, a reporter noticed the resemblance and referred to it as the "plum pudding" model. Thomson hated the name, but it stuck.

result of this gold foil experiment, so Rutherford had inadvertently proven it inaccurate. Rutherford, in turn, proposed his own model, which looks like the one everyone recognizes today (see Figure 9.4). This model has negative electrons orbiting around a positive nucleus. It had two major problems:

1. It didn't explain the black body radiation Planck had modeled with statistics.

2. An orbit is accelerated motion and accelerating charges create light.

The second is a major problem because light carries away energy. This gradually slows down the electrons until they eventually fall into the nucleus. The orbits in Rutherford's model are *not* stable. Why do we still recognize this as the atom? Probably because it was the last time atomic models looked simple.

In 1913, a Danish physicist named Niels Bohr tweaked Rutherford's model trying to fix these problems. He stated that, unlike with gravity, only *some* orbits were possible (see Figure 9.5). Those orbits were given by

$$r_n = \frac{n^2 h^2 \epsilon_0}{\pi Z q^2 m} = \frac{n^2}{Z}\,(0.0529\text{ nm}) \tag{9.1.2}$$

where $n = 1, 2, 3, 4, \ldots$ represents the orbit number and Z is the atomic number (i.e. the number of protons in the nucleus). Unfortunately, if only

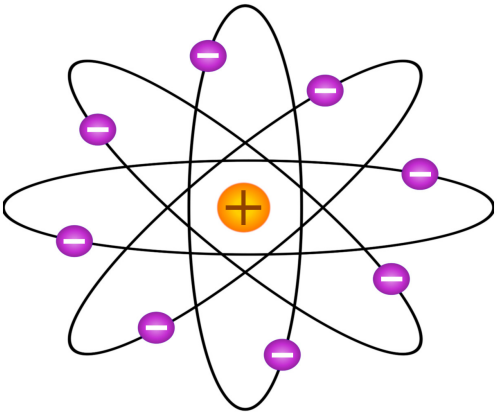


Figure 9.4: This is what Ernest Rutherford envisioned for the atom. With all the positive charge concentrated in a nucleus, the spread in alpha scattering in his gold foil experiment made much more sense.

some orbits are possible, then only *some* energy and angular momentum values are possible. The total energy of an electron is given by

$$E_n = -\frac{Z^2 q^4 m}{8h^2\epsilon_0^2 n^2} = \frac{Z^2}{n^2}\,(-13.6\text{ eV}) \tag{9.1.3}$$

and the angular momentum by

$$L_n = \frac{nh}{2\pi} = n\hbar \tag{9.1.4}$$

where $\hbar = 1.055 \times 10^{-34}$ Js (or 6.582×10^{-16} eVs), pronounced h-bar, is the **reduced Planck's constant**. The energy of the electron is negative because it's in a potential well created by the nucleus (i.e. it's bound to the nucleus).

A transition from a high to a low orbit results in the emission (or loss) of a specific amount of energy, which is given by

$$-\Delta E = E_i - E_f = Z^2\,(13.6\text{ eV})\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right)$$

where n_i is the initial orbit number and n_f is the final orbit number. That energy leaves in the form of a photon, the quantum of light, which has a

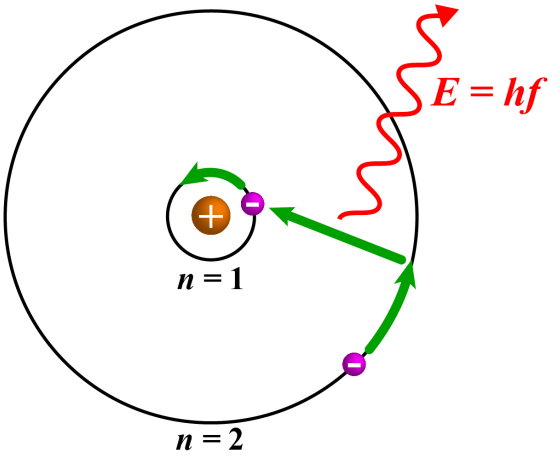


Figure 9.5: Bohr's model of the atom was just a simple tweak of Rutherford's in Figure 9.4. It explains Planck's radiation very well, but only for atoms with a single electron (i.e. Hydrogen, single-ionized Helium, double-ionized Lithium, etc.).

wavelength of

$$hf = \frac{hc}{\lambda} = Z^2\,(13.6\text{ eV})\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right)$$

$$\lambda = \frac{91.13\text{ nm}}{Z^2}\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right)^{-1} \tag{9.1.5}$$

where Z is the atomic number (i.e. the number of protons in the nucleus). Bohr's model explained Planck's radiation model quite well, but the problem of the unstable orbit still remained. The model is also limited because it fails when there is more than one electron in the atom. It was starting to become clear that classical mechanics and electrodynamics were not sufficient to describe what was happening. We needed a new **quantum mechanics**.

Wave-Particle Duality

Two decades into the 20$^{\text{th}}$ century, physicists were still discussing Bohr's model. Not only did it leave some questions unanswered, it also raised a couple more.


Louis de Broglie


Erwin Schrödinger


Max Born


Werner Heisenberg

Figure 9.6: These people were important in the initial development of quantum mechanics.

- How do we resolve the unstable orbit problem?
- Why can electrons only travel along certain orbits?
- Why do our models fail when we try to consider more electrons?
- What makes light so special that it can be both a particle *and* a wave?

These are all questions for which we eventually discovered answers. The first answer we found was to the last question: What makes light so special? The answer: Nothing. It isn't special at all.

In 1924, a french physicist named Louis de Broglie proposed, in his PhD dissertation, that *all* particles can behave as waves. In other words, **wave-particle duality** was not limited to light. He even proposed a way to predict the wavelength and frequency of massive particles like electrons. Using the Planck relation ($E = hf$), the frequency is given by

$$f = \frac{E}{h} = \frac{\gamma m_p c^2}{h}, \tag{9.1.6}$$

where $h = 6.626 \times 10^{-34}$ Js (or 4.136×10^{-15} eVs) is Planck's constant, m_p is the rest mass, $c = 299,792,458$ m/s is the speed of light, and γ is the relativistic gamma factor (Eq. 7.2.9). With wavelength, we have to be more careful. Your first instinct might be to use $v = \lambda f$, but v isn't necessarily the velocity of the particle. Remember, in 1924, we didn't know what these matter waves looked like. Louis de Broglie was in uncharted territory.

We can't make any assumptions in our analysis, so we have to start from what we know about waves. Recall from Eq. 5.5.5 that the general form of

a wave function is

$$y(\vec{r},t) = A\cos\left(-\omega t + \vec{k}\bullet\vec{r} + \varphi_0\right), \tag{9.1.7}$$

where A is the amplitude of the wave, $\vec{r}$ is the position vector, t is time, $\omega = 2\pi f$ is the angular frequency (in radians per second), $k = 2\pi/\lambda$ is the angular wave number (in radians per meter), and φ_0 is a phase angle (in radians). For a massive particle in special relativity (Chapter 7), the scalar product of 4-momentum and 4-position is

$$p^\delta x_\delta = -\gamma E_p t + \gamma\vec{p}_p\bullet\vec{r} = -Et + \vec{p}\bullet\vec{r}$$

and, since $E = hf = \hbar\omega$,

$$p^\delta x_\delta = -\hbar\omega t + \vec{p}\bullet\vec{r} = \hbar\left(-\omega t + \frac{\vec{p}}{\hbar}\bullet\vec{r}\right)$$

which looks a lot like the argument of the cosine in Eq. 9.1.7. By simply matching terms, we can conclude $\vec{p} = \hbar\vec{k}$ with a magnitude of

$$p = \hbar k = \left(\frac{h}{2\pi}\right)\left(\frac{2\pi}{\lambda}\right) = \frac{h}{\lambda}$$

$$\lambda = \frac{h}{p} = \frac{h}{\gamma m_p v}, \tag{9.1.8}$$

where $h = 6.626\times10^{-34}$ Js (or 4.136×10^{-15} eVs) is Planck's constant, m_p is the rest mass, v is the velocity of the particle, and γ is the relativistic gamma factor (Eq. 7.2.9). Eq. 9.1.8 is often called the **de Broglie wavelength**.

You can get a hint at the strange nature of these waves by combining Eqs. 9.1.6 and 9.1.8, which results in

$$v_{\text{phase}} = \lambda f = \frac{h}{p}\frac{E}{h} = \left(\frac{h}{\gamma m_p v}\right)\left(\frac{\gamma m_p c^2}{h}\right)$$

and, since v is the velocity of the particle,

$$v_{\text{phase}} = \frac{E}{p} = \frac{c^2}{v_{\text{particle}}} \tag{9.1.9}$$

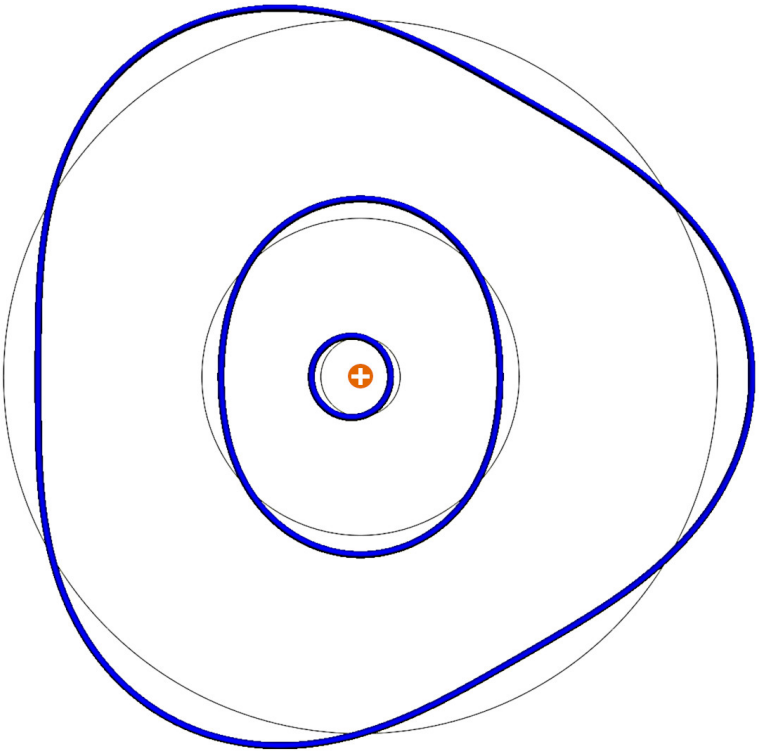


Figure 9.7: This diagrams shows the 3 lowest de Broglie wavelengths for an electron in hydrogen atom. The Bohr orbits (dashed black) are scaled by a factor of n^2 and the waves (solid blue) by a factor of n, which results in each orbit containing n wavelengths. For simplicity, the waves shown are the *group* waves traveling with *group* velocity.

where E is the relativistic energy, p is the relativistic 3-momentum, and $c = 299,792,458$ m/s is the speed of light. The consequence is that $v_{\text{phase}} = v_{\text{particle}}$ only for massless particles because they travel at c. All particles with mass travel with a velocity less than c, so they have a **phase velocity** larger than c. Yes, I said "*larger* than c." It's weird, but information can never be sent this way, so we're not violating any physical laws. The velocity of the particle is usually called a **group velocity** because it's considered a group of multiple waves. We did predict its wavelength by grouping 4-momentum and 4-position, so this shouldn't be too surprising.

Anyway, the idea of matter waves *also* answers the unstable orbit problem. The electrons in an atom can only travel along certain orbits because they're not really orbits at all. They're just the wave paths that result in

constructive interference (i.e. a standing wave). In Figure 9.7, you can see the waves for the lowest three Bohr orbits of the hydrogen atom. Bohr designed his model for non-relativistic particles (i.e. particles traveling with $v \ll c$), which is actually pretty common for quantum mechanics, so we'll say $\gamma \approx 1$. We also know from Newton's second law (Eq. 4.2.6) that

$$\sum \vec{F}_n = m\vec{a}_n \quad \Rightarrow \frac{Zq^2}{4\pi\epsilon_0 r_n^2} = m\frac{v^2}{r_n} \quad \Rightarrow v = \sqrt{\frac{Zq^2}{4\pi\epsilon_0 m r_n}}$$

where r_n is the radius of the Bohr orbit (Eq. 9.1.2). Substituting this into Eq. 9.1.8, we get

$$\lambda = \frac{h}{mv} = \frac{h}{m}\sqrt{\frac{4\pi\epsilon_0 m r_n}{Zq^2}} = \sqrt{\frac{4\pi h^2 \epsilon_0 r_n}{\pi Z q^2 m}}.$$

With a little manipulation, this becomes

$$\lambda = \sqrt{\frac{4\pi^2}{n^2}\left(\frac{n^2 h^2 \epsilon_0}{\pi Z q^2 m}\right) r_n}$$

and the quantity in parentheses is just r_n (Eq. 9.1.2). This drastically simplifies the wavelength of the electron to

$$\lambda = \frac{2\pi r_n}{n} = \frac{n}{Z}\,(0.3324 \text{ nm})\,, \tag{9.1.10}$$

where $2\pi r_n$ is the circumference of the orbit. In other words, you can fit n wavelengths into each Bohr orbit.

The result is important because, if the electron isn't actually orbiting, then it isn't accelerating and it's only going to emit a photon if the matter wave changes. The Bohr orbits are just the natural wavelengths of the electron. Louis de Broglie's solution answered not just one, but three of the questions from the beginning of this section. Unfortunately, it still leaves the issue of generalizing for multiple electrons. It also raises a question we encountered for light in electrodynamics (Section 5.5) and then again in special relativity (Section 7.1): What is *actually* vibrating? The answer is only a bit further down the rabbit hole.

9.2 Waves of Probability

As you can see from the last section, things started to get a bit weird in the 1920s with the discovery that all particles could display wave properties. Unfortunately, this weirdness gets worse. We still needed to figure out what is actually *waving*. The most obvious choice for a particle like an electron was to imagine its energy (and charge) smeared across a Bohr orbit as seen in Figure 9.7. This turned out to be very wrong, but an assumption had to be made to move forward and describe the *behavior* of these particle waves.

Schrödinger's Equation

In Example 5.5.1, we described electromagnetic waves with a **wave function** (Eq. 5.5.5). That wave function, in turn, was a solution to a **wave equation** (Eq. 5.5.2). If we assume the wave function of particle has the same form, then we get

$$\psi(\vec{r},t) = A\cos\left(\vec{k}\bullet\vec{r}-\omega t+\varphi_0\right),$$

which looks a lot like Eq. 9.1.7. The symbol ψ is used to clarify it is a *quantum* wave function. For the sake of generality and ease of use, we'll write this as a complex exponential

$$\psi(\vec{r},t) = A\,e^{i\left(\vec{k}\bullet\vec{r}-\omega t+\varphi_0\right)} = A\,e^{i\varphi_0}\,e^{i\left(\hbar\vec{k}\bullet\vec{r}-\hbar\omega t\right)/\hbar}.$$

Furthermore, since $\vec{p}=\hbar\vec{k}$ and $E=\hbar\omega$,

$$\psi(\vec{r},t) = A\,e^{i(\vec{p}\bullet\vec{r}-Et)/\hbar}, \tag{9.2.1}$$

where $\hbar = 1.055\times10^{-34}$ Js (or 6.582×10^{-16} eVs) and the phase constant $e^{i\varphi_0}$ was merged with the coefficient A.

Contrary to common practice, we've found a general wave function *before* ever finding a wave equation that governs it. We can't conveniently apply these waves to specific cases without a wave equation, which was a problem on Austrian physicist Erwin Schrödinger's mind in 1926. His approach involved the use of energy. The total energy of a particle moving non-relativistically (i.e. $\gamma\approx1$) is given by its **Hamiltonian**,

$$\mathcal{H} = KE + PE = \frac{1}{2}m\vec{v}\bullet\vec{v}+V = \frac{m\vec{v}\bullet m\vec{v}}{2m}+V = \frac{\vec{p}\bullet\vec{p}}{2m}+V, \tag{9.2.2}$$

where $\vec{p} = m\vec{v}$ is the 3-momentum of the particle. We also know this $\mathcal{H}$ and the E from Eq. 9.2.1 should be equivalent, so $\mathcal{H}\psi = E\psi$.

Schrödinger knew this wave equation would need to involve time *and* space derivatives like any other wave equation, so he reverse-engineered it from the wave function (Eq. 9.2.1). The first space derivative of the wave function is

$$\vec{\nabla}\psi = \vec{\nabla}\left(A\, e^{i(\vec{p}\bullet\vec{r}-Et)/\hbar}\right) = \frac{i\vec{p}}{\hbar}\, A\, e^{i(\vec{p}\bullet\vec{r}-Et)/\hbar} = \frac{i\vec{p}}{\hbar}\,\psi,$$

which means

$$\vec{p} = \frac{\hbar}{i}\vec{\nabla} = -i\hbar\vec{\nabla} \tag{9.2.3}$$

and this is where the math starts to get really weird. We saw in Section 3.2 that $\vec{\nabla}$ is an operator, not a quantity. Eq. 9.2.3 implies that $\vec{p}$ is also an operator for the wave function ψ, which will become important in Section 9.3. Consequently, the second space derivative is

$$\vec{\nabla}^2\psi = \vec{\nabla}\bullet\vec{\nabla}\psi = \left(\frac{i\vec{p}}{\hbar}\bullet\frac{i\vec{p}}{\hbar}\right)\psi = \frac{-1}{\hbar^2}\,(\vec{p}\bullet\vec{p})\;\psi,$$

which means

$$\vec{p}\bullet\vec{p} = -\hbar^2\vec{\nabla}^2 \tag{9.2.4}$$

and we have something we can use in Eq. 9.2.2.

The right-hand side of $\mathcal{H}\psi = E\psi$ involves E, which is part of the time term in Eq. 9.2.1. The first time derivative of the wave function is

$$\frac{\partial}{\partial t}\psi = \frac{\partial}{\partial t}\left(A\, e^{i(\vec{p}\bullet\vec{r}-Et)/\hbar}\right) = -\frac{iE}{\hbar}\, A\, e^{i(\vec{p}\bullet\vec{r}-Et)/\hbar} = -\frac{iE}{\hbar}\,\psi$$

$$\Rightarrow E = -\frac{\hbar}{i}\frac{\partial}{\partial t} = i\hbar\frac{\partial}{\partial t}, \tag{9.2.5}$$

which is also an operator. Substituting Eqs. 9.2.2, 9.2.4, and 9.2.5 into $\mathcal{H}\psi = E\psi$, we get

$$\frac{\vec{p}\bullet\vec{p}}{2m}\psi + V\psi = E\psi$$

$$-\frac{\hbar^2}{2m}\vec{\nabla}^2\psi + V\psi = i\hbar\frac{\partial\psi}{\partial t}, \tag{9.2.6}$$

where $\hbar = 1.055 \times 10^{-34}$ Js (or 6.582×10^{-16} eVs), m is the particle's mass, and $i = \sqrt{-1}$ is the imaginary unit. This is called **Schrödinger's equation** and it's the guiding principle of quantum mechanics. If the system is more complicated than a single non-relativistic particle, then we just say

$$\mathcal{H}\psi = i\hbar\frac{\partial\psi}{\partial t}, \tag{9.2.7}$$

where $\mathcal{H}$ is the **Hamiltonian** on the wave function ψ. The form of $\mathcal{H}$ must be determined for the specific case.

As If Things Weren't Crazy Enough...

Eq. 9.2.1 makes no statement of units and, therefore, no statement of what is actually waving. The complex coefficient A (i.e. the amplitude) has whatever units you need. For a mechanical wave on a string, that's meters (m). For electromagnetic waves, that's either newtons per coulomb (N/C) or teslas (T) depending on which field you're discussing. Erwin Schrödinger, like many physicists of his time, just assumed the wave function would measure some already-known property of a particle like position, momentum, mass, charge, etc. Unfortunately, this leads to all sorts of problems.

The de Broglie model shown in Figure 9.7 imagines electrons as smeared out across classical orbits. Let's assume this is true for a moment. Since ψ is complex and measurements like charge are real, we have to be careful. A **complex square** will eliminate the imaginary components, so we'll say

$$\psi^*\psi \equiv \rho \text{ (volumetric charge density)}$$

measured in coulombs per cubic meter ($\mathrm{C/m^3}$). If we subject this charge distribution to an external electric potential, ϕ, then Schrödinger's equation (Eq. 9.2.6) can be written as

$$-\frac{\hbar^2}{2m}\vec{\nabla}^2\psi + (q\phi)\,\psi = i\hbar\frac{\partial\psi}{\partial t}, \tag{9.2.8}$$

where $V = q\phi$ is the electric potential *energy* of the entire electron. Things get more interesting if we add in a magnetic field.

Magnetic fields are weird and their potentials are even weirder. We first saw the magnetic vector potential, $\vec{A}$, in Section 5.6 along with the electric scalar potential, ϕ. However, where $q\phi$ has units of energy, $q\vec{A}$ has units of momentum. We can generalize this in special relativity with qA^δ, where A^δ is the 4-potential (Eq. 7.5.5), since the 4-momentum (Eq. 7.4.22) incorporates energy and momentum. The consequence of this is that 4-momentum is now

$$p^\delta = m_p u^\delta + qA^\delta, \tag{9.2.9}$$

rather than just $m_p u^\delta$. This means conserved non-relativistic 3-momentum is now

$$\vec{p} = m\vec{v} + q\vec{A} \tag{9.2.10}$$

for a charge q with mass m traveling with velocity $\vec{v}$ in a vector potential $\vec{A}$.

Sadly, we used only $m\vec{v}$ in the classical Hamiltonian (Eq. 9.2.2) to derive Schrödinger's equation (Eq. 9.2.6), which means it no longer applies. A quick adjustment results in a kinetic energy term of

$$KE = \frac{1}{2} m\vec{v} \bullet \vec{v} = \frac{m\vec{v} \bullet m\vec{v}}{2m} = \frac{1}{2m}\left(\vec{p} - q\vec{A}\right) \bullet \left(\vec{p} - q\vec{A}\right)$$

and, by Eq. 9.2.3,

$$KE = \frac{1}{2m}\left(-i\hbar\vec{\nabla} - q\vec{A}\right) \bullet \left(-i\hbar\vec{\nabla} - q\vec{A}\right)$$

Therefore, Schrödinger's equation is actually

$$\frac{1}{2m}\left(-i\hbar\vec{\nabla} - q\vec{A}\right) \bullet \left(-i\hbar\vec{\nabla} - q\vec{A}\right)\psi + (q\phi)\,\psi = i\hbar\frac{\partial\psi}{\partial t}, \tag{9.2.11}$$

which applies to non-relativistic charge q (spin $= 0$). If you're dealing specifically with an electron, just say $q = -e = -1.602 \times 10^{-19}$ C.

We have yet to encounter any problems with our original assumption that $\psi^*\psi$ is the volumetric charge density, ρ, because we have yet to ask the right question. The charge distribution of our smeared-out electron is bound to change over time in response to its external influences. This change can be found by

$$\frac{\partial\rho}{\partial t} = \frac{\partial}{\partial t}\left(\psi^*\psi\right) = \psi\frac{\partial\psi^*}{\partial t} + \psi^*\frac{\partial\psi}{\partial t},$$

where we've used the derivative product rule (Eq. 3.1.5). We can use this along with Eq. 5.3.22,

$$\frac{\partial \rho}{\partial t} = -\vec{\nabla} \bullet \vec{J},$$

to find an electric current density, $\vec{J}$. If we can show this current density changes in time, then the charge is accelerating. Since accelerating charges radiate light and we know that shouldn't happen in this circumstance, we'll have a contradiction. The only conclusion will be that our original assumption was false.

We can start by eliminating the time derivatives on the right-hand side using substitutions from Schrödinger's equation (Eq. 9.2.11), which results in

$$\frac{\partial \rho}{\partial t} = -\frac{\psi}{i\hbar}\left[-i\hbar\frac{\partial \psi^*}{\partial t}\right] + \frac{\psi^*}{i\hbar}\left[i\hbar\frac{\partial \psi}{\partial t}\right]$$

$$\begin{aligned}\frac{\partial \rho}{\partial t} &= -\frac{\psi}{i\hbar}\left[\frac{1}{2m}\left(i\hbar\vec{\nabla} - q\vec{A}\right) \bullet \left(i\hbar\vec{\nabla} - q\vec{A}\right)\psi^* + (q\phi)\,\psi^*\right] \\ &\quad + \frac{\psi^*}{i\hbar}\left[\frac{1}{2m}\left(-i\hbar\vec{\nabla} - q\vec{A}\right) \bullet \left(-i\hbar\vec{\nabla} - q\vec{A}\right)\psi + (q\phi)\,\psi\right].\end{aligned}$$

We have to take the complex conjugate of Schrödinger's equation when operating on ψ^*. If we move some factors around,

$$\begin{aligned}\frac{\partial \rho}{\partial t} &= -\frac{1}{2m}\left[\frac{\psi}{i\hbar}\left(i\hbar\vec{\nabla} - q\vec{A}\right) \bullet \left(i\hbar\vec{\nabla} - q\vec{A}\right)\psi^* + \frac{2m}{i\hbar}\psi\,(q\phi)\,\psi^*\right] \\ &\quad -\frac{1}{2m}\left[-\frac{\psi^*}{i\hbar}\left(-i\hbar\vec{\nabla} - q\vec{A}\right) \bullet \left(-i\hbar\vec{\nabla} - q\vec{A}\right)\psi - \frac{2m}{i\hbar}\psi^*\,(q\phi)\,\psi\right]\end{aligned}$$

$$\begin{aligned}\frac{\partial \rho}{\partial t} &= -\frac{1}{2m}\left[\frac{\psi}{i\hbar}\left(i\hbar\vec{\nabla} - q\vec{A}\right) \bullet \left(i\hbar\vec{\nabla} - q\vec{A}\right)\psi^*\right] \\ &\quad -\frac{1}{2m}\left[-\frac{\psi^*}{i\hbar}\left(-i\hbar\vec{\nabla} - q\vec{A}\right) \bullet \left(-i\hbar\vec{\nabla} - q\vec{A}\right)\psi\right],\end{aligned}$$

and expand the two binomial products,

$$\begin{aligned}\frac{\partial \rho}{\partial t} &= -\frac{1}{2m}\left[\frac{\psi}{i\hbar}\left(i^2\hbar^2\vec{\nabla}^2 - q\vec{A} \bullet i\hbar\vec{\nabla} - i\hbar\vec{\nabla} \bullet q\vec{A} + q^2\vec{A} \bullet \vec{A}\right)\psi^*\right] \\ &\quad -\frac{1}{2m}\left[-\frac{\psi^*}{i\hbar}\left(i^2\hbar^2\vec{\nabla}^2 + q\vec{A} \bullet i\hbar\vec{\nabla} + i\hbar\vec{\nabla} \bullet q\vec{A} + q^2\vec{A} \bullet \vec{A}\right)\psi\right]\end{aligned}$$

$$\begin{aligned}\frac{\partial\rho}{\partial t} &= -\frac{1}{2m}\left[i\hbar\psi\vec{\nabla}^2\psi^* - \psi q\vec{A}\bullet\vec{\nabla}\psi^* - \psi\vec{\nabla}\bullet q\vec{A}\psi^* + \frac{q^2}{i\hbar}\psi\vec{A}\bullet\vec{A}\psi^*\right] \\ &\quad -\frac{1}{2m}\left[-i\hbar\psi^*\vec{\nabla}^2\psi - \psi^* q\vec{A}\bullet\vec{\nabla}\psi - \psi^*\vec{\nabla}\bullet q\vec{A}\psi - \frac{q^2}{i\hbar}\psi^*\vec{A}\bullet\vec{A}\psi\right],\end{aligned}$$

a total of four terms cancel giving us

$$\begin{aligned}\frac{\partial\rho}{\partial t} &= -\frac{1}{2m}\left[i\hbar\psi\vec{\nabla}^2\psi^* - \psi q\vec{A}\bullet\vec{\nabla}\psi^* - \psi\vec{\nabla}\bullet q\vec{A}\psi^*\right] \\ &\quad -\frac{1}{2m}\left[-i\hbar\psi^*\vec{\nabla}^2\psi - \psi^* q\vec{A}\bullet\vec{\nabla}\psi - \psi^*\vec{\nabla}\bullet q\vec{A}\psi\right].\end{aligned}$$

We need to be a bit more careful with the remaining terms. If we do some *voodoo math* by adding in two opposite extra terms ($i\hbar\vec{\nabla}\psi\vec{\nabla}\psi^*$), we get

$$\begin{aligned}\frac{\partial\rho}{\partial t} &= -\frac{1}{2m}\left[i\hbar\psi\vec{\nabla}^2\psi^* + i\hbar\vec{\nabla}\psi\vec{\nabla}\psi^* - \psi q\vec{A}\bullet\vec{\nabla}\psi^* - \psi\vec{\nabla}\bullet q\vec{A}\psi^*\right] \\ &\quad -\frac{1}{2m}\left[-i\hbar\psi^*\vec{\nabla}^2\psi - i\hbar\vec{\nabla}\psi^*\vec{\nabla}\psi - \psi^* q\vec{A}\bullet\vec{\nabla}\psi - \psi^*\vec{\nabla}\bullet q\vec{A}\psi\right].\end{aligned}$$

Regrouping a few things and using the derivative product rule (Eq. 3.1.5) in reverse results in

$$\begin{aligned}\frac{\partial\rho}{\partial t} &= -\frac{1}{2m}\left[i\hbar\left(\psi\vec{\nabla}^2\psi^* + \vec{\nabla}\psi\vec{\nabla}\psi^*\right) - \left(\psi^* q\vec{A}\bullet\vec{\nabla}\psi + \psi\vec{\nabla}\bullet q\vec{A}\psi^*\right)\right] \\ &\quad -\frac{1}{2m}\left[-i\hbar\left(\psi^*\vec{\nabla}^2\psi + \vec{\nabla}\psi^*\vec{\nabla}\psi\right) - \left(\psi q\vec{A}\bullet\vec{\nabla}\psi^* + \psi^*\vec{\nabla}\bullet q\vec{A}\psi\right)\right]\end{aligned}$$

$$\begin{aligned}\frac{\partial\rho}{\partial t} &= -\frac{1}{2m}\left[i\hbar\vec{\nabla}\bullet\left(\psi\vec{\nabla}\psi^*\right) - \vec{\nabla}\bullet\left(\psi q\vec{A}\psi^*\right)\right] \\ &\quad -\frac{1}{2m}\left[-i\hbar\vec{\nabla}\bullet\left(\psi^*\vec{\nabla}\psi\right) - \vec{\nabla}\bullet\left(\psi^* q\vec{A}\psi\right)\right]\end{aligned}$$

$$\frac{\partial\rho}{\partial t} = -\vec{\nabla}\bullet\frac{1}{2m}\left[\psi\left(i\hbar\vec{\nabla} - q\vec{A}\right)\psi^* + \psi^*\left(-i\hbar\vec{\nabla} - q\vec{A}\right)\psi\right], \qquad (9.2.12)$$

which is now written as a divergence. It is clear from Eqs. 9.2.12 and 5.3.22 that the current density is given by

$$\vec{J} = \frac{1}{2m}\left[\psi\left(i\hbar\vec{\nabla} - q\vec{A}\right)\psi^* + \psi^*\left(-i\hbar\vec{\nabla} - q\vec{A}\right)\psi\right]. \qquad (9.2.13)$$

Sometimes this is written as

$$\vec{J} = \frac{1}{2m}\left[i\hbar\left(\psi\vec{\nabla}\psi^* - \psi^*\vec{\nabla}\psi\right) - q\vec{A}\psi^*\psi\right] \tag{9.2.14}$$

or, using the definition of the momentum operator (Eq. 9.2.3),

$$\vec{J} = \frac{1}{2m}\left[(\psi^*\vec{p}\,\psi - \psi\,\vec{p}\,\psi^*) - q\vec{A}\psi^*\psi\right] \tag{9.2.15}$$

to more clearly separate the magnetic contribution.

This electric current density is dependent on the wave function ψ. According to the general Schrödinger's equation (Eq. 9.2.7), ψ has a non-zero time derivative if it has a non-zero Hamiltonian. This implies the current density, $\vec{J}$, will also have a non-zero time derivative meaning the charge distribution will radiate light. Therefore,

$$\psi^*\psi \neq \rho \text{ (volumetric charge density)},$$

and we are back where we started. What kind of waves are these? Later in 1926, a German physicist named Max Born suggested an answer: waves of *probability*. The idea is the electron isn't actually smeared-out across a Bohr orbit. As Richard Feynman once said, "The electron is either here, or there, or somewhere else; but, wherever it is, it is a point charge."

The quantity $\psi^*\psi$ is just the **probability density** (generally, probability per unit volume) of finding the electron in any particular place. Another way to say this is $\psi^*\psi\,dx\,dy\,dz$ is the probability of finding the electron in the infinitesimal cube between (x, y, z) and $(x + dx, y + dy, z + dz)$. In a more practical sense,

$$P = \int_{x_1}^{x_2}\int_{y_1}^{y_2}\int_{z_1}^{z_2}\psi^*\psi\,dx\,dy\,dz \tag{9.2.16}$$

is the probability of finding the electron from x_1 to x_2 along x, from y_1 to y_2 along y, *and* from z_1 to z_2 along z. For consistency with the concept of probability, we always say the probability of finding the electron *somewhere* is

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\psi^*\psi\,dx\,dy\,dz = 1 \tag{9.2.17}$$

or 100%. This is called the **normalization condition**.

Assuming $\psi^*\psi$ is a probability density, that makes $\vec{J}$ the **probability current density** (i.e. the rate of flow of probability per unit area). If that's the case, then Eq. 9.2.12 can be interpreted as a local conservation of probability. In other words, if the probability increases at one position, then it must decrease at another position. Furthermore, the probability must have "flowed" between those two points. This gives us a decent foundation from which to interpret *any* quantum measurement we might take.

9.3 Quantum Measurements

Measuring anything on a quantum level can be tricky at best and impossible at worst. However, the purpose of a theoretical treatment is not to take measurements, but to make predictions. In Section 9.2, we showed the wave function, ψ, represents only the probability of a particular measurement. We cannot make any predictions about what that measurement *will* be, only what it *might* be. This is a big obstacle, but that doesn't mean we should just give up. We can still get plenty of information. It just wont be as much as you'd like.

Observables vs. States

First, we need to determine what it is we're trying to predict/measure. Whatever it is, we call it an **observable**. It can really be anything, but some common examples are

- position, $\vec{r}$,
- momentum, $\vec{p}$,
- energy, $\mathcal{H}$,
- angular momentum, $\vec{L}$, and
- spin, $\vec{S}$.

In quantum mechanics, each of these is represented by an **operator** (a tool first introduced in Section 2.1). We briefly saw this behavior during the

derivation of Schrödinger's equation (Eq. 9.2.6). In general, the observables are functions of position, momentum, and time.

The wave function, $\psi(\vec{r}, t)$, is often called a **state** because it represents the quantum state of the particle. It contains all the information we can ever really predict about a particle. When we use observables as operators on these states, we can make predictions about that particular observable. The most common prediction is the **expectation value**,

$$\langle Q \rangle = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \psi^* Q \psi \, dx \, dy \, dz \,, \tag{9.3.1}$$

of an arbitrary observable, $Q(\vec{r}, \vec{p}, t)$. If you were to perform many measurements of the observable, Q, on a particle in the state, ψ, then Eq. 9.3.1 predicts the average value of those measurements. Specifically, this is a *weighted* average just like the atomic masses on the periodic table. For average atomic mass, there is a finite number of possibilities (i.e. it's **discrete**), so

$$\langle m \rangle = \sum_i m_i \, (\text{isotope abundance})_i = \sum_i m_i P_i,$$

where the "isotope abundance" is a fraction between 0 and 1 (i.e. a percent between 0 and 100%). If you look at a single atom in a random sample of an element, the abundance is just the probability of seeing that particular isotope. However, an observable like position is **continuous**, so the sum includes an infinite number of possibilities. We would write this as

$$\langle x \rangle = \int x \, dP = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x \, \rho \, dx \, dy \, dz$$

where ρ is the probability density.

In order to guarantee Eq. 9.3.1 always results in a *real* value (as opposed to a complex value), Q must be a **Hermitian** operator. Mathematically, that is

$$\iiint_{\text{all space}} \psi^* \, (Q\psi) \; dx \, dy \, dz = \iiint_{\text{all space}} (Q\psi)^* \, \psi \, dx \, dy \, dz, \tag{9.3.2}$$

where both sides result in the same expectation value for Q (and we've used "all space" in place of the cumbersome $\pm\infty$ limits). We can and will keep

using integrals and functions to make quantum predictions, but they're not always the best way. Linear algebra says that functions can also be expressed as vectors or matrices, which is sometimes more convenient. It's always nice to have alternative options.

Bra-Ket Notation

Representing states as functions was certainly favored by Erwin Schrödinger. However, by the 1930s, it was becoming clear that this method had its drawbacks. Max Born and Werner Heisenberg had already begun using matrices. In 1939, Paul Dirac published an article titled *A New Notation For Quantum Mechanics* in which he attempted to bridge the gap between these different methods.

Dirac's method involved representing states as vectors in a **Hilbert space**. A Hilbert space is just a space with more than three dimensions like the one David Hilbert used for spacetime in general relativity (see Chapter 8). In quantum mechanics though, the "space" is not necessarily spatial and can have as many "dimensions" as necessary. Oh, and it's also complex.

The new notation can defined by extending the brackets of the expectation value (Eq. 9.3.1) as

$$\langle Q \rangle = \langle \psi | \, Q \, | \psi \rangle \,, \tag{9.3.3}$$

where $|\psi\rangle$ is the **state** vector (replacing the state function) for the particle. The $|\psi\rangle$ is called a "ket" vector and the $\langle\psi|$ is called a "bra" vector, which is a play on words since we often refer to this notation as "bracket" (**bra-ket**) notation. The bra and ket vectors must be complex conjugates,

$$\langle \psi | = |\psi\rangle^* \,, \tag{9.3.4}$$

and the expectation value of Q must be real, so

$$\langle \psi | \, Q \, | \psi \rangle = (\langle \psi | \, Q \, | \psi \rangle)^* = (Q \, | \psi \rangle)^* \, (\langle \psi |)^* = (\langle \psi | \, Q^*) \, (|\psi\rangle)$$

$$\langle \psi | \, Q \, | \psi \rangle = \langle \psi | \, Q^* \, | \psi \rangle \,, \tag{9.3.5}$$

which means Q is still a **Hermitian** operator.

Bra-ket notation can be a bit abstract, but it has its advantages over the other methods. Schrödinger couldn't write state *functions* without first

deciding whether they would be in terms of position or momentum. This is formally called choosing a mathematical **basis** (first seen at the end of Section 1.3 and then thoroughly defined in Section 6.4). A basis is just a set of vectors for which all other vectors are a linear combination. Whenever possible, we choose an **orthonormal basis**, which is just a set of basis vectors that have a length of one (i.e. unit vectors) and are all orthogonal (e.g. $\{\hat{x}, \hat{y}, \hat{z}\}$ is an orthonormal position basis).

Writing states as abstract vectors frees us from this burden. State vectors are still normalized like state functions, so

$$\langle\psi|\,|\psi\rangle = 1, \tag{9.3.6}$$

but what exactly does an operation like this mean? Let's start by trying to show how the methods compare. According to Eq. 9.2.17,

$$\iiint_{\text{all space}} \psi^*(\vec{r},t)\,\psi(\vec{r},t)\;dx\,dy\,dz = 1,$$

where ψ is written in the position basis. Well, $|\psi\rangle$ is a vector, which is only *projected* onto a basis. We can project onto the Cartesian position basis using

$$|\psi\rangle = \iiint_{\text{all space}} |\vec{r}\rangle\,\langle\vec{r}|\,|\psi\rangle\;dx\,dy\,dz\,. \tag{9.3.7}$$

Combining Eqs. 9.3.6 and 9.3.7, we get

$$\langle\psi| \iiint_{\text{all space}} |\vec{r}\rangle\,\langle\vec{r}|\,|\psi\rangle\;dx\,dy\,dz = 1$$

$$\iiint_{\text{all space}} \langle\psi|\,|\vec{r}\rangle\,\langle\vec{r}|\,|\psi\rangle\;dx\,dy\,dz = 1,$$

which matches Eq. 9.2.17 as long as $\psi(\vec{r},t) = \langle\vec{r}|\,|\psi\rangle$ and $\psi^*(\vec{r},t) = \langle\psi|\,|\vec{r}\rangle$.

We can, therefore, interpret $\langle\psi|\,|\psi\rangle$ as a **probability density**. It then follows that

$$(\langle\psi|\,|\psi\rangle)^*\,\langle\psi|\,|\psi\rangle = \|\langle\psi|\,|\psi\rangle\|^2 = 1$$

is a probability of 100%, which makes sense since both states in the operation are the same. As a less trivial circumstance, consider a particle in a state $|\psi\rangle$. The **probability** of finding it in a state $|\phi\rangle$ would be

$$P = (\langle\phi|\,|\psi\rangle)^*\,\langle\phi|\,|\psi\rangle = \|\langle\phi|\,|\psi\rangle\|^2\,, \tag{9.3.8}$$

which is only 100% if $|\psi\rangle = |\phi\rangle$. The advantage here is we never had to project onto any particular basis to discover the probability.

Many undergraduate quantum textbooks will focus on functions and integrals because they're far more familiar to students. As a result, those students tend to be at a disadvantage when taking graduate courses or reading articles on their own. Undergraduate courses that also expose students to Max Born's matrix method put those students in a slightly better position. However, matrix operations can still be more complicated (e.g. multiplication involves a transpose matrix) than bra-ket operations, which may cause some confusion. I will do my best to expose you to all three methods (functions, matrices, and bra-ket) as we go.

Time-Independent Schrödinger Equation

Let's say we want to solve Schrödinger equation (Eq. 9.2.7) for $\psi(\vec{r}, t)$. You might be wondering: How do we go about doing that? Isn't there an infinite number of possible solutions and, therefore, an infinite number of possible states? Well, yes. The best way to solve a partial differential equation like this is to find all the separable solutions,

$$\psi(\vec{r}, t) = \Psi(\vec{r})\,U(t)\,, \tag{9.3.9}$$

where Ψ is only a function of position and U is only a function of time. All *general* solutions will then be linear combinations of those solutions,

$$\psi(\vec{r}, t) = \sum_{n=1}^{\infty} c_n \Psi_n(\vec{r})\,U_n(t)\,, \tag{9.3.10}$$

where c_n are just constant complex coefficients.

If we apply the separation of variables (Eq. 9.3.9) to Schrödinger equation (Eq. 9.2.7), then we get

$$\mathcal{H}\,(\Psi U) = i\hbar \frac{\partial\,(\Psi U)}{\partial t}.$$

Since $\mathcal{H}$ only has space-derivatives and the right-hand side only has time derivatives,

$$U\mathcal{H}\Psi = i\hbar\Psi\frac{\partial U}{\partial t}$$

$$\frac{1}{\Psi}\mathcal{H}\Psi = i\hbar\left(\frac{1}{U}\frac{\partial U}{\partial t}\right), \tag{9.3.11}$$

and we can't cancel the Ψ's on the left because $\mathcal{H}$ must operate first. The benefit here is that everything on the left is only a function of space and everything on the right is only a function of time. We also know that space and time are independent variables, so changing one will *not necessarily* change the other. The only case in which this doesn't contradict Eq. 9.3.11 is when both sides are constant.

The easy part of solving Eq. 9.3.11 is the time-dependence. Since $\mathcal{H}$ has units of energy, it seems fitting to call the stuff on the right E,

$$E = i\hbar\left(\frac{1}{U}\frac{\partial U}{\partial t}\right),$$

which is very similar to Eq. 9.2.5 with the exception that this E is a real constant (rather than an operator). If we move some things around, then we get

$$\frac{\partial U}{\partial t} = \frac{E}{i\hbar}U = \frac{-iE}{\hbar}U.$$

There is only one function with a first derivative proportional to itself: the natural exponential function. Therefore,

$$U(t) = e^{-iEt/\hbar}, \tag{9.3.12}$$

which is sometimes called the **time-evolution factor** because it governs how the wave function, $\psi(\vec{r},t)$, changes in time. Judging from its form, it rotates the state in the complex plane. This factor results in separable solutions with the form

$$\psi(\vec{r},t) = \Psi(\vec{r})\,e^{-iEt/\hbar}, \tag{9.3.13}$$

where we now only need to solve for $\Psi(\vec{r})$ (the space dependence).

The separable solutions given by Eq. 9.3.13 are called **stationary states** because the particle won't transition out of these states on its own. They are *stable* states. If we substitute them into Schrödinger equation (Eq. 9.2.7), then

$$\mathcal{H}\left(\Psi e^{-iEt/\hbar}\right) = i\hbar\frac{\partial}{\partial t}\left(\Psi e^{-iEt/\hbar}\right)$$

$$e^{-iEt/\hbar}\mathcal{H}\Psi = Ee^{-iEt/\hbar}\Psi$$

$$\boxed{\mathcal{H}\Psi = E\Psi}\,. \tag{9.3.14}$$

We call this the **time-independent Schrödinger equation** because its solutions, $\Psi(\vec{r})$, are independent of time. Sometimes these solutions are also referred to as states, which can get really confusing. They are not states! State functions are given by Eq. 9.3.13 and include the time-dependence.

You can also write Eq. 9.3.14 in bra-ket notation as long as the partial-state is a ket vector, $|\Psi\rangle$. It takes the form

$$\boxed{\mathcal{H}\,|\Psi\rangle = E\,|\Psi\rangle}\,. \tag{9.3.15}$$

Essentially, what this means is that, if you were to take a measurement of $\mathcal{H}$ of a particle in the partial-state Ψ, you would get a definite energy of E and the partial-state of the particle would not change as a result. If the particle were in some other full-state,

$$|\psi(\vec{r},t)\rangle = e^{-iEt/\hbar}\,|\Psi(\vec{r})\rangle\,, \tag{9.3.16}$$

then the equation might look something like

$$\mathcal{H}\,|\psi\rangle = E\,|\phi\rangle\,,$$

where it has switched to a ϕ full-state as a result of the measurement.

Saying "partial-state" is already getting annoying, so let's borrow another name from mathematics. For equations like Eq. 9.3.14, mathematicians call the solutions "eigenfunctions" ("eigen" is german word for "own" or "inherent"). If we adopt that formalism, then Ψ would be an **eigenfunction** and $|\Psi\rangle$ would be an **eigenvector**. You could call either an **eigenstate** to

distinguish it from a full-state. This also makes E an **eigenvalue** of Eq. 9.3.15.

Bra-ket notation is extremely advantageous when using the time-independent Schrödinger equation (Eq. 9.3.15) because the eigenstates, being vectors, will all be orthonormal. Mathematically, we say

$$\langle \Psi_i | \, | \Psi_j \rangle = \delta_{ij}, \tag{9.3.17}$$

where δ_{ij} is the Kronecker delta (Eq. 6.2.2). This is useful because it's a *complete* set of orthonormal vectors. That makes it an orthonormal *basis* for the Hilbert space, meaning all possible full-states can be written as a linear combination of those eigenvectors:

$$|\psi\rangle = \sum_{n=1}^{\infty} c_n e^{-iE_n t/\hbar} \, |\Psi_n\rangle \, , \tag{9.3.18}$$

where c_n are just constant complex coefficients. We already stated this for functions as Eq. 9.3.10. Using Eq. 9.3.12, we get

$$\psi(\vec{r}, t) = \sum_{n=1}^{\infty} c_n \Psi_n(\vec{r}) \, e^{-iE_n t/\hbar} \, , \tag{9.3.19}$$

but it is much clearer why this is true using vectors.

Heisenberg Uncertainty Principle

Being limited only to probabilities isn't our only problem. In 1927, Werner Heisenberg suggested that some observables were incompatible with one another. His paper *Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik* (German for "*About the intuitive content of quantum theoretical kinematics and mechanics*"), he stated this problem for only two pairs of observables:

- position, $\vec{r}$, and momentum, $\vec{p}$
- energy, $\mathcal{H}$, and state lifetime, Δt.

The state lifetime is how long a particle stays in a particular state whose Hamiltonian is $\mathcal{H}$.

All of these observables are operators, so the question is: "Are those operations commutative?" As it turns out, position and momentum do not (i.e. $\vec{r} \bullet \vec{p}\,\psi \neq \vec{p} \bullet \vec{r}\,\psi$). Their actual relationship is given by something called a **commutator**,

$$[\vec{r}, \vec{p}] = \vec{r} \bullet \vec{p} - \vec{p} \bullet \vec{r}, \tag{9.3.20}$$

which is an operator itself. To find its non-zero value, we need to first make it operate on a general state ψ:

$$[\vec{r}, \vec{p}]\,\psi = \vec{r} \bullet \vec{p}\,\psi - \vec{p} \bullet \vec{r}\,\psi.$$

The momentum operator, $\vec{p}$, is given by Eq. 9.2.3, so

$$\begin{aligned} [\vec{r}, \vec{p}]\,\psi &= \vec{r} \bullet \left(-i\hbar\vec{\nabla}\right)\psi - \left(-i\hbar\vec{\nabla}\right) \bullet \vec{r}\,\psi \\ &= i\hbar\left[-\vec{r} \bullet \vec{\nabla}\psi + \vec{\nabla} \bullet (\vec{r}\,\psi)\right]. \end{aligned}$$

Using the chain rule for derivatives (Eq. 3.1.2)on the last term, we get

$$\begin{aligned} [\vec{r}, \vec{p}]\,\psi &= i\hbar\left[-\vec{r} \bullet \vec{\nabla}\psi + \left(\vec{\nabla} \bullet \vec{r}\right)\,\psi + \vec{r} \bullet \vec{\nabla}\psi\right] \\ &= i\hbar\left[\left(\vec{\nabla} \bullet \vec{r}\right)\,\psi\right]. \end{aligned}$$

However, a quick use of Eqs. 1.1.1 and 3.2.1, results in

$$\vec{\nabla} \bullet \vec{r} = \frac{\partial x}{\partial x} + \frac{\partial y}{\partial y} + \frac{\partial z}{\partial z} = 1 + 1 + 1 = 3$$

so

$$[\vec{r}, \vec{p}] = 3i\hbar. \tag{9.3.21}$$

Traditionally, this is written in only one dimension. By the vector dot product (Eq. 2.2.2), we can say

$$[\vec{r}, \vec{p}] = [x, p_x] + [y, p_y] + [z, p_z] = 3i\hbar.$$

Since the orientation of the axes doesn't matter, all three terms should be equal, so

$$\boxed{[x, p_x] = i\hbar}. \tag{9.3.22}$$

This is called the **canonical commutation relation** and it was a major result in Heisenberg's paper.

The next question that's probably on most of your minds: Why does any of this even matter? This incompatibility has direct consequences on how *precise* our predictions are allowed to be. Let's consider a new statistical quantity: the **standard deviation**, σ (see Example 9.4.5 for context). It measures how much variation the predicted measurements will have around the expectation value and is given by

$$\sigma_A = \sqrt{\left\langle (A - \langle A \rangle)^2 \right\rangle} = \sqrt{\langle A^2 \rangle - \langle A \rangle^2} \tag{9.3.23}$$

for some arbitrary observable A. The square eliminates unimportant negative signs and the square root allows the units to match those of A.

In 1928, a German mathematician named Hermann Weyl applied Eq. 9.3.23 to quantum predictions. We'll start by taking the square of the standard deviation (i.e. the variance),

$$\sigma_A^2 = \left\langle (A - \langle A \rangle)^2 \right\rangle, \tag{9.3.24}$$

and applying Eq. 9.3.3, which results in

$$\begin{aligned} \sigma_A^2 &= \langle \psi | \left(A - \langle A \rangle\right)^2 | \psi \rangle \\ &= \langle \left(A - \langle A \rangle\right) \psi | \left(A - \langle A \rangle\right) \psi \rangle \end{aligned}$$

for some arbitrary observable A. Borrowing the Cauchy-Schwarz inequality from mathematics,

$$\langle a | a \rangle \langle b | b \rangle \geq \| \langle a | b \rangle \|^2, \tag{9.3.25}$$

we can say

$$\sigma_A^2 \sigma_B^2 \geq \| \langle \left(A - \langle A \rangle\right) \psi | \left(B - \langle B \rangle\right) \psi \rangle \|^2 \tag{9.3.26}$$

as long as $|a\rangle = |\left(A - \langle A \rangle\right) \psi\rangle$ and $|b\rangle = |\left(B - \langle B \rangle\right) \psi\rangle$. We can simplify the expectation value on the right-hand side a bit by expanding it

$$\begin{aligned} & \langle \left(A - \langle A \rangle\right) \psi | \left(B - \langle B \rangle\right) \psi \rangle \\ = \; & \langle \psi | \left(A - \langle A \rangle\right) \left(B - \langle B \rangle\right) | \psi \rangle \\ = \; & \langle \psi | AB - A \langle B \rangle - \langle A \rangle B + \langle A \rangle \langle B \rangle | \psi \rangle \\ = \; & \langle \psi | AB | \psi \rangle - \langle \psi | A | \psi \rangle \langle B \rangle - \langle A \rangle \langle \psi | B | \psi \rangle + \langle A \rangle \langle B \rangle \langle \psi | \psi \rangle \end{aligned}$$

and, by Eqs. 9.3.3 and 9.3.6,

$$= \langle AB \rangle - \langle A \rangle \langle B \rangle - \langle A \rangle \langle B \rangle + \langle A \rangle \langle B \rangle$$

$$\langle (A - \langle A \rangle) \psi | \, | (B - \langle B \rangle) \psi \rangle = \langle AB \rangle - \langle A \rangle \langle B \rangle \,. \tag{9.3.27}$$

This turns Eq. 9.3.26 into

$$\sigma_A^2 \sigma_B^2 \geq \| \langle AB \rangle - \langle A \rangle \langle B \rangle \|^2 \tag{9.3.28}$$

for two arbitrary observables A and B.

Unfortunately, Eq. 9.3.28 isn't very useful. Since it still depends on the expectation values of A and B, it depends on the specific experiment being done. However, if we can somehow relate this to their commutator,

$$[A, B] \equiv AB - BA \,, \tag{9.3.29}$$

then we'll have something dependent only on what A and B *are* rather their specific expectation values. The *complex* square on the right-hand side of Eq. 9.3.28 is necessary because $\langle AB \rangle - \langle A \rangle \langle B \rangle$ isn't necessarily *real*. Any complex number, z, will always obey

$$\begin{aligned} \|z\|^2 &= \|\mathrm{Re}(z) + i \, \mathrm{Im}(z)\|^2 \\ &= [\mathrm{Re}(z) + i \, \mathrm{Im}(z)] \, [\mathrm{Re}(z) - i \, \mathrm{Im}(z)] \\ &= \mathrm{Re}(z)^2 + \mathrm{Im}(z)^2 \,. \end{aligned}$$

where $\mathrm{Re}(z)$ denotes the real part of z and $\mathrm{Im}(z)$ denotes the imaginary part of z. From the definition of z, we also know

$$\begin{aligned} z - z^* &= [\mathrm{Re}(z) + i \, \mathrm{Im}(z)] - [\mathrm{Re}(z) - i \, \mathrm{Im}(z)] \\ &= \mathrm{Re}(z) + i \, \mathrm{Im}(z) - \mathrm{Re}(z) + i \, \mathrm{Im}(z) \\ &= 2i \, \mathrm{Im}(z) \end{aligned}$$

$$\mathrm{Im}(z) = \frac{1}{2i} (z - z^*) \,. \tag{9.3.30}$$

Since, with squares, everything is now positive and real, we can say

$$\begin{aligned} \mathrm{Re}(z)^2 + \mathrm{Im}(z)^2 &\geq \mathrm{Im}(z)^2 \\ \|z\|^2 &\geq \left[\frac{1}{2i}(z - z^*)\right]^2 . \end{aligned}$$

Using this in Eq. 9.3.28 with $z = \langle AB\rangle - \langle A\rangle\langle B\rangle$ and $z^* = \langle BA\rangle - \langle B\rangle\langle A\rangle$, we get

$$\begin{aligned} \sigma_A^2\sigma_B^2 &\geq \|z\|^2 \geq \left[\frac{1}{2i}(z - z^*)\right]^2 \\ \sigma_A^2\sigma_B^2 &\geq \left[\frac{1}{2i}\left(\langle AB\rangle - \langle A\rangle\langle B\rangle - \langle BA\rangle + \langle B\rangle\langle A\rangle\right)\right]^2 \\ \sigma_A^2\sigma_B^2 &\geq \left[\frac{1}{2i}\left(\langle AB\rangle - \langle BA\rangle\right)\right]^2 \\ \sigma_A^2\sigma_B^2 &\geq \left[\frac{1}{2i}\langle AB - BA\rangle\right]^2 , \end{aligned}$$

which just involves the commutator (Eq. 9.3.29). The final result is then

$$\sigma_A^2\sigma_B^2 \geq \left[\frac{1}{2i}\langle [A,B]\rangle\right]^2 , \tag{9.3.31}$$

for two arbitrary observables A and B.

Eq. 9.3.31 often referred to as the **Heisenberg uncertainty principle**, though Heisenberg never derived anything this general. If A and B are *compatible* observables, then $[A,B] = 0$ and $\sigma_A^2\sigma_B^2 \geq 0$. This means either standard deviation (σ_A or σ_B) could easily be as close to zero as you need it to be. Some notable pairs of compatible observables are

- Energy and Total Angular Momentum ($L^2 \equiv L_x^2 + L_y^2 + L_z^2$):

$$\left[\mathcal{H}, L^2\right] = 0 \tag{9.3.32}$$

- Energy and Angular Momentum along z:

$$[\mathcal{H}, L_z] = 0 \tag{9.3.33}$$

- Total Angular Momentum and Angular Momentum along z:

$$\left[L^2, L_z\right] = 0 \tag{9.3.34}$$

which will become very important in Chapter 10. If A and B are *incompatible* observables, then $[A, B] \neq 0$ and you'll have to find Eq. 9.3.31 for that specific pair. For example, the components of angular momentum are incompatible with each other as shown by

- Angular Momentum along x and Angular Momentum along y:

$$[L_x, L_y] = i\hbar L_z \tag{9.3.35}$$

- Angular Momentum along y and Angular Momentum along z:

$$[L_y, L_z] = i\hbar L_x \tag{9.3.36}$$

- Angular Momentum along z and Angular Momentum along x:

$$[L_z, L_x] = i\hbar L_y \tag{9.3.37}$$

or, more succinctly,

$$[L_i, L_j] = i\hbar\varepsilon_{ijk} L_k \tag{9.3.38}$$

where ε_{ijk} is the Levi-Civita pseudotensor (Eq. 6.6.4).

In the case of one-dimensional position and momentum (Eq. 9.3.22), the uncertainty principle reduces to

$$\begin{aligned} \sigma_x^2 \sigma_{p_x}^2 &\geq \left[\frac{1}{2i}\langle i\hbar\rangle\right]^2 \\ \sigma_x^2 \sigma_{p_x}^2 &\geq \left[\frac{\hbar}{2}\right]^2 \end{aligned}$$

or, written in a more traditional way,

$$\sigma_x \sigma_{p_x} \geq \frac{\hbar}{2}. \tag{9.3.39}$$

This is the result Heisenberg had arrived at in his 1927 paper and it has two consequences.

1. Predicted measurements of position and momentum will always vary around the expectation value.

2. The more precise you can predict position, the less precise your prediction of momentum will be (and vice versa).

It should be emphasized that these precision issues are *not* due to limits of our technology. Eq. 9.3.31 was derived using only generic statistics and the knowledge that matter behaves like waves. It is a fundamental result of the mechanics of matter waves and, therefore, a fundamental property of the universe.

Recall that, for any observable, there exists a set of stationary states (Eq. 9.3.13) that are described partially by eigenstates. These stationary states are states of *definite* value. If observables are compatible, then they share a complete set of eigenstates (i.e. it is possible to find a particle in a stationary state of *both* observables at the same time). That means both can be predicted with precision. However, if observables are incompatible, then you will *never* find the particle in a stationary state of both at the same time. That means if one has a definite value, then the other does not (i.e. it is less precise).

9.4 Simple Models

Quantum mechanics is a very broad field with many models for many different situations. It's hardly something anyone could cover *completely* in a whole book, let alone just a chapter or two. However, there are a few models:

1. Infinite Square Wells,

2. Finite Square Wells, and

3. Harmonic Oscillators;

that are important from an educational standpoint because they're simple. They give those new to the subject a framework from which to understand how Schrödinger's equation (Eq. 9.2.6) is used. They may not be *realistic*, but may be useful in some extreme circumstances. We'll save the more realistic models for Chapter 10.

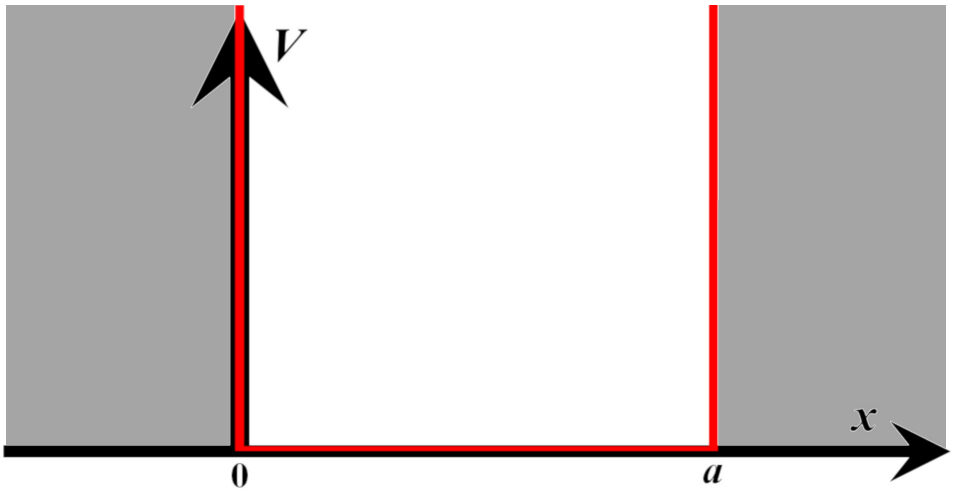


Figure 9.8: This shows the (one-dimensional) infinite square well's potential energy (Eq. 9.4.1) graphed against position, x.

Infinite Square Well

An infinite square well (or just infinite well) is the simplest potential energy to provide a particle. In one dimension, it's usually stated as

$$\boxed{V(x) = \begin{cases} 0, & \text{if } 0 \leq x \leq a \\ \infty, & \text{otherwise} \end{cases}} \tag{9.4.1}$$

for a well with an arbitrary width, a. Figure 9.8 shows clearly why we refer to this as a "well." There is no potential energy inside, but it's infinite outside. Basically, this particle is never getting out because the it's in an infinitely deep hole. In three dimensions, it's usually stated as

$$V(x,y,z) = \begin{cases} 0, & \text{if } 0 \leq x \leq a_x, 0 \leq y \leq a_y, \text{ and } 0 \leq z \leq a_z \\ \infty, & \text{otherwise} \end{cases} \tag{9.4.2}$$

for a well with that is a_x by a_y by a_z.

This model isn't ordinarily very accurate. *Approximately* though, neutron stars and white dwarfs are prime candidates! They're *nearly* inescapable and the pressure is so high that nuclei stop having individual electron clouds. In my master's thesis, I used Eq. 9.4.2 as the quantum potential energy for electrons in a white dwarf star and it was surprisingly accurate. Black holes

might *actually* be inescapable, but they've hypothetically compressed the matter into a singularity, so the "well" wouldn't have size which makes this model useless.

Example 9.4.1

What are the stationary states (and corresponding energies) for a non-relativistic particle in a one-dimensional infinite square well?

- First, there are no stationary states (i.e. $\psi(x,t) = 0$) outside the well. It is impossible to achieve infinite potential energy. All we really need to find are the stationary states inside the well.
- Inside the well, the time-independent Schrödinger equation (Eq. 9.3.14) will take the form

$$\mathcal{H}\Psi = -\frac{\hbar^2}{2m}\vec{\nabla}^2\Psi = E\Psi$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} = E\Psi \tag{9.4.3}$$

  where we've set $V = 0$ and $\vec{\nabla}^2$ is only in one dimension, x.
- Moving some things around, we get

$$\frac{\partial^2\Psi}{\partial x^2} = -\frac{2mE}{\hbar^2}\Psi,$$

  which is a very common differential equation. There are only two functions with second derivatives proportional to the negative of themselves: $\sin(kx)$ and $\cos(kx)$. The general solution will be a linear combination of the two,

$$\Psi(x) = C_1\sin(kx) + C_2\cos(kx)\,,$$

  where C_1 and C_2 are just constants. The k, another constant, is determined by the differential equation

$$\frac{\partial^2\Psi}{\partial x^2} = -k^2\Psi = -\frac{2mE}{\hbar^2}\Psi$$

$$\Rightarrow k = \frac{\sqrt{2mE}}{\hbar},$$

so the eigenstates take the form

$$\Psi(x) = C_1 \sin\left(\frac{\sqrt{2mE}}{\hbar}x\right) + C_2 \cos\left(\frac{\sqrt{2mE}}{\hbar}x\right).$$

Now we just need to determine C_1 and C_2.

- Note: **All eigenstates must be continuous (and finite) over all space and must have a first derivative that is continuous (and finite) over all space.** This is because its second derivative must exist in all space due to Schrödinger equation.

- Coefficients are usually determined by initial conditions or, in this case, boundary conditions. We know a little something about the behavior of $\Psi(x)$ at the boundaries: $x = 0$ and $x = a$. Since the particle cannot exist outside of the well, we can say

$$\Psi(0) = \Psi(a) = 0 \tag{9.4.4}$$

so the eigenstates remain continuous. Using the condition at $x = 0$, we get

$$\Psi(0) = 0 = C_1 \sin(0) + C_2 \cos(0) = 0 + C_2 = C_2,$$

meaning the cosine term disappears. The solution has reduced to

$$\Psi(x) = C_1 \sin\left(\frac{\sqrt{2mE}}{\hbar}x\right).$$

- Using the condition at $x = a$, we get

$$\Psi(a) = 0 = C_1 \sin\left(\frac{\sqrt{2mE}}{\hbar}a\right).$$

We could also set $C_1 = 0$, but that would be trivial and completely useless, so

$$\sin\left(\frac{\sqrt{2mE}}{\hbar}a\right) = 0.$$

This means, based on the behavior of sine,

$$\frac{\sqrt{2mE}}{\hbar}a = \pm n\pi,$$

where n is a whole number (i.e. $n = 0, 1, 2, 3, \ldots$). Unfortunately, $n = 0$ just results in $\Psi(x) = 0$ again, so not useful. Also, the $\pm$ doesn't really tell us anything since $\sin(-x) = -\sin(x)$ and the negative will just get absorbed into C_1. Therefore, we'll say

$$\frac{\sqrt{2mE}}{\hbar}a = n\pi,$$

where n is a whole number (i.e. $n = 1, 2, 3, \ldots$)

- Each value of n corresponds to a definite energy level, E_n, for the particle. Solving for E gives us

$$\frac{2mE}{\hbar^2}a^2 = n^2\pi^2$$

$$\boxed{E_n = \frac{n^2\pi^2\hbar^2}{2ma^2}}, \tag{9.4.5}$$

where n is a whole number (i.e. $n = 1, 2, 3, \ldots$).

- However, we still haven't determined the value of C_1. Remember the normalization condition (Eq. 9.2.17)? If we intend on interpreting this stationary state as a probability wave, then the probability of finding the particle *somewhere* should be 1 (i.e. 100%). This is why we call this final constant the **normalization constant**. Using Eq. 9.2.17 and the fact that

$$\psi^*\psi = \left(\Psi^* e^{iEt/\hbar}\right)\left(\Psi e^{-iEt/\hbar}\right) = \Psi^*\Psi,$$

we get

$$\int_{-\infty}^{+\infty} \Psi^*\Psi \, dx = 1.$$

Since Ψ is entirely real in this example, but is different over different values of x,

$$\int_{-\infty}^{0} \Psi^2 \, dx + \int_{0}^{a} \Psi^2 \, dx + \int_{a}^{+\infty} \Psi^2 \, dx = 1$$

$$\int_{-\infty}^{0} 0 \, dx + \int_{0}^{a} \left[C_1 \sin\left(\frac{\sqrt{2mE}}{\hbar} x \right) \right]^2 dx + \int_{a}^{+\infty} 0 \, dx = 1$$

$$C_1^2 \int_{0}^{a} \sin^2\left(\frac{\sqrt{2mE}}{\hbar} x \right) dx = 1.$$

Since we know from mathematics that

$$\int_a^b \sin^2(kx) \, dx = \int_a^b \tfrac{1-\cos(2kx)}{2} dx = \frac{x}{2} - \frac{\sin(2kx)}{4k} \Big|_a^b, \qquad (9.4.6)$$

we can say

$$C_1^2 \left[\frac{x}{2} - \frac{\hbar}{4\sqrt{2mE}} \sin\left(2\frac{\sqrt{2mE}}{\hbar} x \right) \right]_0^a = 1$$

$$C_1^2 \left[\frac{a}{2} - 0 - 0 + 0 \right] = 1 \quad \Rightarrow C_1 = \sqrt{\frac{2}{a}}.$$

Therefore the eigenstates take the form

$$\Psi_n(x) = \sqrt{\frac{2}{a}} \sin\left(\frac{\sqrt{2mE_n}}{\hbar} x \right)$$

or, better yet,

$$\boxed{\Psi_n(x) = \sqrt{\frac{2}{a}} \sin\left(\frac{n\pi}{a} x \right)}, \qquad (9.4.7)$$

where n is a whole number (i.e. $n = 1, 2, 3, \ldots$). The first three are shown in Figure 9.9.

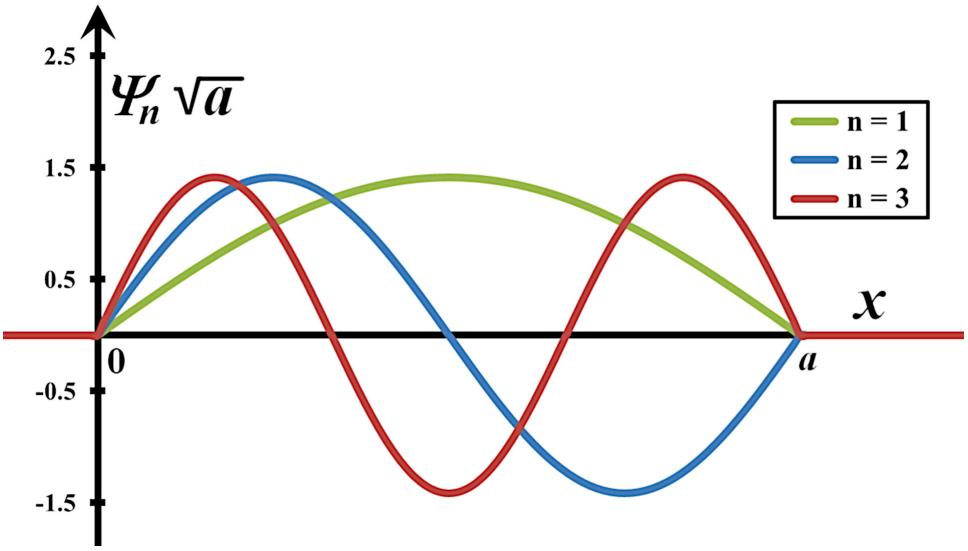


Figure 9.9: This shows the first three eigenstates of the (one-dimensional) infinite square well given by Eq. 9.4.7.

- Unfortunately, the eigenstates are only the stationary states at $t = 0$. In general, stationary states are given by Eq. 9.3.13, so

$$\psi_n(x,t) = \Psi_n(x)\, e^{-iEt/\hbar}$$

$$\boxed{\psi_n(x,t) = \sqrt{\frac{2}{a}} \sin\!\left(\frac{n\pi}{a}x\right) e^{-in^2\pi^2\hbar\, t/\left(2ma^2\right)}}, \tag{9.4.8}$$

  where n is a whole number (i.e. $n = 1, 2, 3, \ldots$). As we mentioned when defining the time-evolution factor (Eq. 9.3.12), these stationary states are the eigenstates rotating about the x-axis (see Figure 9.10). That rotation will occur at a faster rate for higher energies.

- The particle in this well doesn't necessarily have to be in one of the stationary states. According to Eq. 9.3.19, all the possible solutions of the time-*dependent* Schrödinger equation (Eq. 9.2.7) take the form

$$\psi(x,t) = \sum_{n=1}^{\infty} c_n \sqrt{\frac{2}{a}} \sin\!\left(\frac{n\pi}{a}x\right) e^{-in^2\pi^2\hbar\, t/\left(2ma^2\right)}, \tag{9.4.9}$$

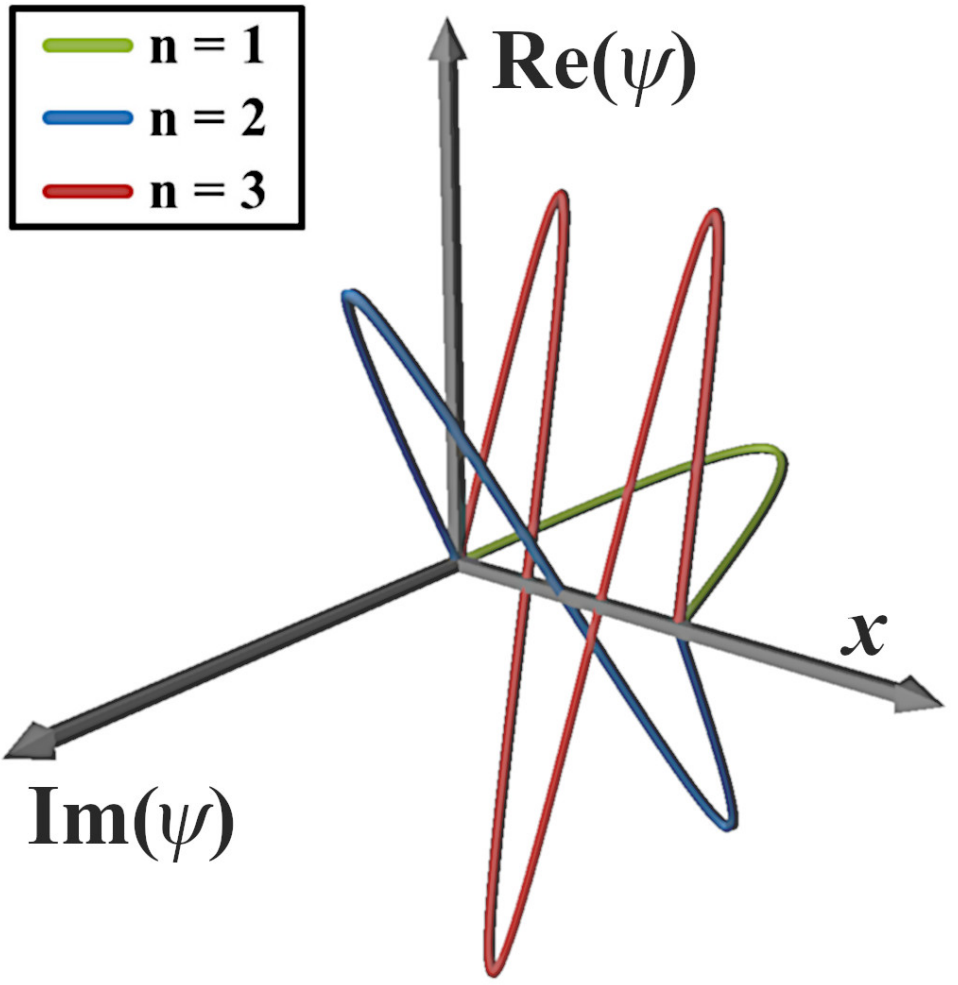


Figure 9.10: This shows the first three stationary states of the (one-dimensional) infinite square well given by Eq. 9.4.8 at some $t \neq 0$. For these stationary states at $t = 0$, refer to Figure 9.9.

where n is a whole number (i.e. $n = 1, 2, 3, \ldots$) and c_n are just constant coefficients. Any particle in *any* state (in this well) will have a wave function that looks like Eq. 9.4.9.

Example 9.4.2

What are the stationary states (and corresponding energies) for a non-relativistic particle in a three-dimensional infinite square well?

- Just like with the one-dimensional case, there are no stationary states (i.e. $\psi(x,t) = 0$) outside the well. It is impossible to achieve infinite potential energy. All we really need to find are the stationary states inside the well.

- Inside the well, the time-independent Schrödinger equation (Eq. 9.3.14) will take the form

$$\mathcal{H}\Psi = -\frac{\hbar^2}{2m}\vec{\nabla}^2\Psi = E\Psi$$

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2}\right) = E\Psi \tag{9.4.10}$$

  where we've set $V = 0$ and $\vec{\nabla}^2$ has been expanded into three dimensions.

- The Hamiltonian is made of commuting parts,

$$\left[\frac{\partial^2}{\partial x^2}, \frac{\partial^2}{\partial y^2}\right] = \left[\frac{\partial^2}{\partial y^2}, \frac{\partial^2}{\partial z^2}\right] = \left[\frac{\partial^2}{\partial z^2}, \frac{\partial^2}{\partial x^2}\right] = 0,$$

  because partial derivatives are *always* commutative. The mathematical consequence is the eigenstates are separable,

$$\Psi(x,y,z) = X(x)\,Y(y)\,Z(z)\,. \tag{9.4.11}$$

  This is not always the case, since different potential energy functions can cause problems. It only worked this time because $V = 0$.

- If we plug Eq. 9.4.11 into Eq. 9.4.10, then

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2(XYZ)}{\partial x^2}+\frac{\partial^2(XYZ)}{\partial y^2}+\frac{\partial^2(XYZ)}{\partial z^2}\right)=E\,XYZ$$

$$-\frac{\hbar^2}{2m}\left(YZ\frac{\partial^2 X}{\partial x^2}+XZ\frac{\partial^2 Y}{\partial y^2}+XY\frac{\partial^2 Z}{\partial z^2}\right)=E\,XYZ.$$

Dividing through by XYZ, we get

$$-\frac{\hbar^2}{2m}\left(\frac{1}{X}\frac{\partial^2 X}{\partial x^2}+\frac{1}{Y}\frac{\partial^2 Y}{\partial y^2}+\frac{1}{Z}\frac{\partial^2 Z}{\partial z^2}\right)=E$$

$$-\frac{\hbar^2}{2m}\frac{1}{X}\frac{\partial^2 X}{\partial x^2}-\frac{\hbar^2}{2m}\frac{1}{Y}\frac{\partial^2 Y}{\partial y^2}-\frac{\hbar^2}{2m}\frac{1}{Z}\frac{\partial^2 Z}{\partial z^2}=E.$$

We can see the three individual terms add to be a constant. However, since the terms are not codependent (i.e. they're functions of different independent variables), then their sum can be constant *only* if each term is individually constant. Therefore, this is just three independent differential equations:

$$-\frac{\hbar^2}{2m}\frac{1}{X}\frac{\partial^2 X}{\partial x^2}=E_x,\quad -\frac{\hbar^2}{2m}\frac{1}{Y}\frac{\partial^2 Y}{\partial y^2}=E_y,\quad -\frac{\hbar^2}{2m}\frac{1}{Z}\frac{\partial^2 Z}{\partial z^2}=E_z;$$

where $E=E_x+E_y+E_z$.

- These three differential equations look identical to the one-dimensional case (Eq. 9.4.3), so they have the same general solutions. Since the boundary conditions are also identical, the specific solutions will be the same. Using these variables, we get eigenstates in the form

$$X(x)=\sqrt{\frac{2}{a_x}}\sin\left(\frac{n_x\pi}{a_x}x\right) \tag{9.4.12a}$$

$$Y(y)=\sqrt{\frac{2}{a_y}}\sin\left(\frac{n_y\pi}{a_y}y\right) \tag{9.4.12b}$$

$$Z(z)=\sqrt{\frac{2}{a_z}}\sin\left(\frac{n_z\pi}{a_z}z\right) \tag{9.4.12c}$$

with energies

$$E_{n_x} = \frac{n_x^2 \pi^2 \hbar^2}{2ma_x^2} \tag{9.4.13a}$$

$$E_{n_y} = \frac{n_y^2 \pi^2 \hbar^2}{2ma_y^2} \tag{9.4.13b}$$

$$E_{n_y} = \frac{n_z^2 \pi^2 \hbar^2}{2ma_z^2} \tag{9.4.13c}$$

where the appropriate quantities have directional labels.

- If we piece Eq. Sets 9.4.12 and 9.4.13 together, then the eigenstates are

$$\boxed{\Psi_{n_x n_y n_z} = \sqrt{\frac{8}{a_x a_y a_z}} \sin\left(\frac{n_x \pi}{a_x} x\right) \sin\left(\frac{n_y \pi}{a_y} y\right) \sin\left(\frac{n_z \pi}{a_z} z\right)} \tag{9.4.14}$$

with energies

$$E_{n_x n_y n_z} = \frac{n_x^2 \pi^2 \hbar^2}{2ma_x^2} + \frac{n_y^2 \pi^2 \hbar^2}{2ma_y^2} + \frac{n_z^2 \pi^2 \hbar^2}{2ma_z^2}$$

$$\boxed{E_{n_x n_y n_z} = \frac{\pi^2 \hbar^2}{2m} \left(\frac{n_x^2}{a_x^2} + \frac{n_y^2}{a_y^2} + \frac{n_z^2}{a_z^2} \right)}, \tag{9.4.15}$$

maintaining directional labels. Similar to Eq. 9.3.13, the full stationary states can be found by

$$\psi_{n_x n_y n_z}(x, y, z, t) = \Psi_{n_x n_y n_z}(x, y, z)\ e^{-iE_{n_x n_y n_z} t/\hbar}, \tag{9.4.16}$$

with some really nasty subscripts.

- There is something interesting about the three-dimensional case though. If $a_x = a_y = a_z = a$, then different states can share the same energy. For example, even though the corresponding state functions are different,

$$E_{211} = E_{121} = E_{112} = \frac{\pi^2 \hbar^2}{2m} \left(\frac{2^2}{a^2} + \frac{1^2}{a^2} + \frac{1^2}{a^2} \right) = \frac{3\pi^2 \hbar^2}{ma^2}$$

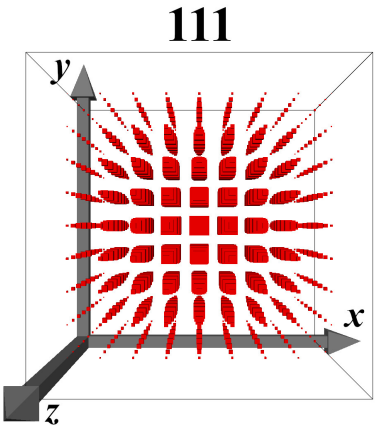


Figure 9.11: This is a representation of the lowest energy level ($n_x n_y n_z$ is 111) for the three-dimensional infinite well with $a_x = a_y = a_z = a$ (boundaries shown). The value of $\Psi(x, y, z)$ (Eq. 9.4.14) is shown as the size of the blocks. The color red indicates the values of Ψ are positive.

are all the same energy (shown in Figure 9.12). When this happens, we say those states are **degenerate**. The phenomenon of **degeneracy** is a consequence of working in a three-dimensional world and is common in more complex examples as well.

Finite Square Well

Most sources of potential energy aren't even close to infinite, so finite square wells (or just finite wells) are a little more realistic. We could start by swapping out the infinity in Eq. 9.4.1 for an arbitrary value V_0,

$$V(x) = \begin{cases} 0, & \text{if } 0 \leq x \leq a \\ V_0, & \text{otherwise} \end{cases}$$

for a well with an arbitrary width, a. We can do better though. If we're calling this a "well," then the potential energy outside (i.e. in normal space) should be zero and inside should be negative (i.e. a deficit). We'll say

$$V(x) = \begin{cases} -V_0, & \text{if } 0 \leq x \leq a \\ 0, & \text{otherwise} \end{cases}$$

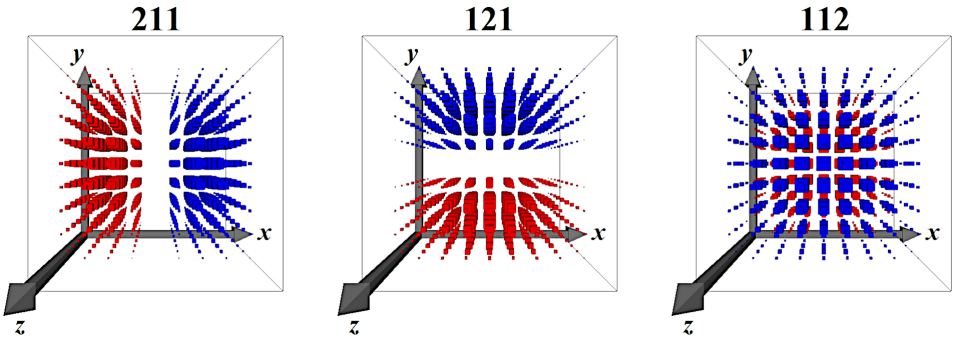


Figure 9.12: This is a representation of the second energy level for the three-dimensional infinite well with $a_x = a_y = a_z = a$ (boundaries shown). Each diagram is labeled by $n_x n_y n_z$. The value of $\Psi(x, y, z)$ (Eq. 9.4.14) is shown as the size of the blocks. The color red indicates the values of Ψ are positive and the color blue indicates the values of Ψ are negative.

for a well with an arbitrary width, a. This is much more realistic, but it doesn't make our lives very easy.

In Chapter 1, we said the coordinate system was just a tool and some choices are better than others. If we shift the coordinate to the middle of the well, then

$$\boxed{V(x) = \begin{cases} -V_0, & \text{if } -\frac{a}{2} \le x \le +\frac{a}{2} \\ 0, & \text{otherwise} \end{cases}} \tag{9.4.17}$$

for a well with an arbitrary width, a (shown in Figure 9.13). This potential energy function has **symmetry**, so the solutions will also have symmetry. In this case, it's symmetry over the vertical axis (i.e. the potential energy function is *even*), so it has the same value when you transform by $x \to -x$.

If we do the same transformation on the time-independent Schrödinger equation (Eq. 9.3.14), we get

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi(-x)}{\partial x^2} + V(-x)\ \Psi(-x) = E\,\Psi(-x)\,.$$

However, since $V(-x) = V(x)$, this becomes

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi(-x)}{\partial x^2} + V(x)\ \Psi(-x) = E\,\Psi(-x)\,,$$

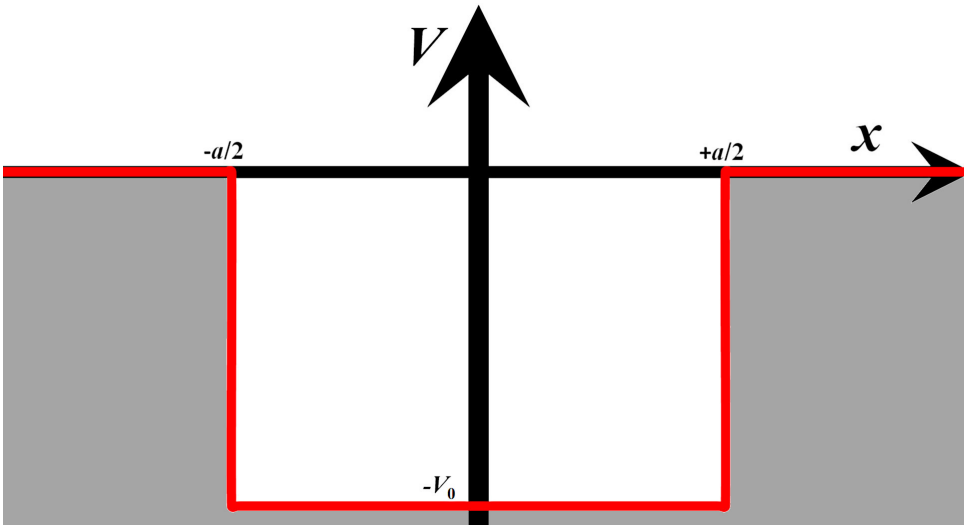


Figure 9.13: This shows the (one-dimensional) finite square well's potential energy (Eq. 9.4.17) graphed against position, x.

which is just the time-independent Schrödinger equation (Eq. 9.3.14) applied to $\Psi(-x)$. If $\Psi(x)$ and $\Psi(-x)$ are *both* general solutions to the *same* differential equation, then they can only differ by a constant multiple, C. We also know the solution must eventually be normalized, so the only options are

$$\Psi(x) = C\,\Psi(-x) = \pm\Psi(-x)\,, \tag{9.4.18}$$

positive for *even* (vertical axis symmetry) and negative for *odd* (origin symmetry). No other eigenstates are possible because they would violate Eq. 9.4.18.

Example 9.4.3

An electron, moving at non-relativistic speeds, is in a one-dimensional finite square well ($V_0 = 250$ eV, $a = 0.1$ nm). What are the stationary states (and corresponding energies) for the electron?

- We're going to solve this problem by staying as general as possible for as long as possible. This will allow us to see results common to all finite wells.
- Unlike the infinite well (see Example 9.4.1), it *is* possible to find the particle beyond the walls (i.e. for $|x| > a/2$). Think of it like probability

sort of "bleeding" through the walls. You'd expect the probability to drop off quickly, but it's still possible. Since the potential energy there is zero, the time-independent Schrödinger equation (Eq. 9.3.14) will take the form

$$\mathcal{H}\Psi = -\frac{\hbar^2}{2m}\vec{\nabla}^2\Psi = E\Psi$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} = E\Psi, \tag{9.4.19}$$

where $\vec{\nabla}^2$ is only in one dimension, x.

- Moving some things around, we get

$$\frac{\partial^2\Psi}{\partial x^2} = \frac{-2mE}{\hbar^2}\Psi,$$

which looks a lot like what we got for the infinite well. However, remember E is negative because it's an energy deficit, so

$$\alpha = \frac{\sqrt{-2mE}}{\hbar} \tag{9.4.20}$$

must be positive and real. There is only one function with a second derivative proportional to the positive of itself: the exponential, $e^{\pm\alpha x}$. Therefore, the eigenstates are a linear combination of the two,

$$\Psi(x) = C_1 e^{\sqrt{-2mE}\,x/\hbar} + C_2 e^{-\sqrt{-2mE}\,x/\hbar} \tag{9.4.21}$$

for the regions where $|x| > a/2$.

- Inside the well (i.e. for $|x| \leq a/2$), the potential energy is $-V_0$, so the time-independent Schrödinger equation (Eq. 9.3.14) will take the form

$$\mathcal{H}\Psi = -\frac{\hbar^2}{2m}\vec{\nabla}^2\Psi + (-V_0)\,\Psi = E\Psi$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} - V_0\Psi = E\Psi, \tag{9.4.22}$$

where $\vec{\nabla}^2$ is only in one dimension, x.

- Moving some things around, we get

$$\frac{\partial^2 \Psi}{\partial x^2} = \frac{-2m\,(E+V_0)}{\hbar^2}\Psi.$$

For this particle to be affected by the well (i.e. in a **bound state**), the particle's energy E must be somewhere between zero and $-V_0$, so $E+V_0$ must be positive and

$$k = \frac{\sqrt{2m\,(E+V_0)}}{\hbar} \tag{9.4.23}$$

must be positive and real (just like in the infinite well). There are only two functions with second derivatives proportional to the negative of themselves: $\sin(kx)$ and $\cos(kx)$. Therefore, the eigenstates are a linear combination of the two,

$$\Psi(x) = C_3 \sin\left(\frac{\sqrt{2m(E+V_0)}}{\hbar}x\right) + C_4 \cos\left(\frac{\sqrt{2m(E+V_0)}}{\hbar}x\right) \tag{9.4.24}$$

for the region where $|x| \leq a/2$.

- The complete solution would be written piecewise like the potential energy function (Eq. 9.4.17), so we have

$$\Psi = \begin{cases} C_1 e^{\sqrt{-2mE}\,x/\hbar} + C_2 e^{-\sqrt{-2mE}\,x/\hbar} & , \text{ if } x < -\dfrac{a}{2} \\ C_3 \sin\left(\dfrac{\sqrt{2m(E+V_0)}}{\hbar}x\right) + C_4 \cos\left(\dfrac{\sqrt{2m(E+V_0)}}{\hbar}x\right) & , \text{ if } |x| \leq \dfrac{a}{2} \\ C_5 e^{\sqrt{-2mE}\,x/\hbar} + C_6 e^{-\sqrt{-2mE}\,x/\hbar} & , \text{ if } x > +\dfrac{a}{2} \end{cases}$$

for each of the three regions. Now we just need to use boundary conditions to solve for the constant coefficients.

- Note: **All eigenstates must be continuous (and finite) over all space and must have a first derivative that is continuous (and finite) over all space.** This is because its second derivative must exist in all space due to Schrödinger equation.

- First, we'll start with the infinite boundaries (i.e. $x \to \pm\infty$). As we approach negative infinity,

$$e^{\sqrt{-2mE}\,x/\hbar} \to 0 \text{ and } e^{-\sqrt{-2mE}\,x/\hbar} \to \infty,$$

so the only conclusion is $C_2 = 0$ since Ψ must be finite. As we approach positive infinity,

$$e^{\sqrt{-2mE}\,x/\hbar} \to \infty \text{ and } e^{-\sqrt{-2mE}\,x/\hbar} \to 0,$$

so the only conclusion is $C_5 = 0$ since Ψ must be finite. Therefore, the eigenstates reduce to

$$\Psi = \begin{cases} C_1 e^{\sqrt{-2mE}\,x/\hbar} & , \text{ if } x < -\dfrac{a}{2} \\ C_3 \sin\left(\dfrac{\sqrt{2m(E+V_0)}}{\hbar}x\right) + C_4 \cos\left(\dfrac{\sqrt{2m(E+V_0)}}{\hbar}x\right) & , \text{ if } |x| \leq \dfrac{a}{2} \\ C_6 e^{-\sqrt{-2mE}\,x/\hbar} & , \text{ if } x > +\dfrac{a}{2} \end{cases}$$

for each of the three regions. The boundaries at the walls of the well are a bit trickier. We'll need to reduce these functions a little more before we can apply them.

- Recalling Eq. 9.4.18, we know the solutions must be either *even* or *odd*, so we'll never use the sine and cosine at the same time. It also means $C_6 = \pm C_1$, positive for even and negative for odd, because the functions should "mirror" each other. Therefore, the even eigenstates are in the form

$$\boxed{\Psi_{\text{even}} = \begin{cases} C_1 e^{\sqrt{-2mE}\,x/\hbar} & , \text{ if } x < -\dfrac{a}{2} \\ C_4 \cos\left(\dfrac{\sqrt{2m(E+V_0)}}{\hbar}x\right) & , \text{ if } |x| \leq \dfrac{a}{2} \\ C_1 e^{-\sqrt{-2mE}\,x/\hbar} & , \text{ if } x > +\dfrac{a}{2} \end{cases}} \tag{9.4.25}$$

and the odd eigenstates are in the form

$$\boxed{\Psi_{\text{odd}} = \begin{cases} C_1 e^{\sqrt{-2mE}\,x/\hbar} & , \text{ if } x < -\dfrac{a}{2} \\ C_3 \sin\left(\dfrac{\sqrt{2m(E+V_0)}}{\hbar} x\right) & , \text{ if } |x| \leq \dfrac{a}{2} \\ -C_1 e^{-\sqrt{-2mE}\,x/\hbar} & , \text{ if } x > +\dfrac{a}{2} \end{cases}} \tag{9.4.26}$$

which still apply to *all* finite square wells described by Eq. 9.4.17.

- The next task is to find C_1, C_3, and C_4. Before we can normalize the wave function, we need there to only be one coefficient per solution. This is achieved through a boundary condition at $x = a/2$. Since all eigenstates must be continuous in all space, C_1 is

$$\Psi_{\text{even}}\big|_{x=\frac{a}{2}} = C_1 e^{-\sqrt{-2mE}\,a/(2\hbar)} = C_4 \cos\left(\frac{\sqrt{2m(E+V_0)}}{\hbar}\frac{a}{2}\right)$$

$$\Rightarrow \boxed{C_{1,\text{even}} = C_4\, e^{\sqrt{-2mE}\,a/(2\hbar)}\, \cos\left(\frac{\sqrt{2m(E+V_0)}}{\hbar}\frac{a}{2}\right)} \tag{9.4.27}$$

for the even solutions and

$$\Psi_{\text{odd}}\big|_{x=\frac{a}{2}} = -C_1 e^{-\sqrt{-2mE}\,a/(2\hbar)} = C_3 \sin\left(\frac{\sqrt{2m(E+V_0)}}{\hbar}\frac{a}{2}\right)$$

$$\Rightarrow \boxed{C_{1,\text{odd}} = -C_3 e^{\sqrt{-2mE}\,a/(2\hbar)} \sin\left(\frac{\sqrt{2m(E+V_0)}}{\hbar}\frac{a}{2}\right)} \tag{9.4.28}$$

for the odd solutions.

- According to the normalization condition (Eq. 9.2.17), the probability of finding the particle *somewhere* should be 1 (i.e. 100%), so

$$\int_{-\infty}^{+\infty} \psi^*\psi\, dx = \int_{-\infty}^{+\infty} \left(\Psi^* e^{iEt/\hbar}\right)\left(\Psi e^{-iEt/\hbar}\right)\, dx = \int_{-\infty}^{+\infty} \Psi^*\Psi\, dx = 1.$$

Since Ψ is entirely real in this example, but is different over different values of x,

$$\int_{-\infty}^{-a/2} \Psi^2 \, dx + \int_{-a/2}^{+a/2} \Psi^2 \, dx + \int_{+a/2}^{+\infty} \Psi^2 \, dx = 1.$$

However, we know the parts outside the finite well "mirror" each other, so

$$2\int_{-\infty}^{-a/2} \Psi^2 \, dx + \int_{-a/2}^{+a/2} \Psi^2 \, dx = 1, \tag{9.4.29}$$

where we've doubled the first term to make up for the loss of the third term.

- We'll start with the even solutions, but using α (Eq. 9.4.20) and k (Eq. 9.4.23) will make for an easier read. Eq. 9.4.29 becomes

$$2\int_{-\infty}^{-a/2} \left[C_1 e^{\alpha x}\right]^2 \, dx + \int_{-a/2}^{+a/2} \left[C_4 \cos(kx)\right]^2 \, dx = 1,$$

where we've substituted from Eq. 9.4.25. If we use Eq. 9.4.27, then

$$2\int_{-\infty}^{-a/2} \left[C_4 e^{\alpha a/2} \cos\left(\tfrac{ka}{2}\right) e^{\alpha x}\right]^2 \, dx + \int_{-a/2}^{+a/2} \left[C_4 \cos(kx)\right]^2 \, dx = 1$$

$$2C_4^2 e^{\alpha a} \cos^2\left(\tfrac{ka}{2}\right) \int_{-\infty}^{-a/2} e^{2\alpha x} \, dx + C_4^2 \int_{-a/2}^{+a/2} \cos^2(kx) \, dx = 1.$$

Since we know from mathematics that

$$\int_a^b e^{\beta x} \, dx = \left.\frac{e^{\beta x}}{\beta}\right|_a^b \tag{9.4.30}$$

and

$$\int_a^b \cos^2(kx) \, dx = \int_a^b \tfrac{1+\cos(2kx)}{2} dx = \left.\frac{x}{2} + \frac{\sin(2kx)}{4k}\right|_a^b, \tag{9.4.31}$$

we can say

$$2C_4^2 e^{\alpha a}\cos^2\left(\tfrac{ka}{2}\right)\left[\frac{e^{2\alpha x}}{2\alpha}\right]_{-\infty}^{-a/2} + C_4^2\left[\frac{x}{2}+\frac{\sin(2kx)}{4k}\right]_{-a/2}^{+a/2} = 1$$

$$2C_4^2 e^{\alpha a}\cos^2\left(\tfrac{ka}{2}\right)\left[\frac{e^{-\alpha a}}{2\alpha}-0\right] + C_4^2\left[2\left(\frac{a}{4}+\frac{\sin(ka)}{4k}\right)\right] = 1$$

$$C_4^2\left[\frac{1}{\alpha}\cos^2\left(\tfrac{ka}{2}\right)+\frac{a}{2}+\frac{1}{2k}\sin(ka)\right] = 1$$

$$\boxed{C_4 = \sqrt{\frac{1}{\frac{1}{\alpha}\cos^2\left(\frac{ka}{2}\right)+\frac{a}{2}+\frac{1}{2k}\sin(ka)}}}, \tag{9.4.32}$$

where α is given by Eq. 9.4.20) and k by Eq. 9.4.23. The $a/2$ term is the same as in the *infinite* well, but the additional terms (because they involve α and k) depend on the energy of the state. The consequence is C_4 is not universal for the *finite* well like it was for the infinite case.

- Going through a nearly identical process, C_3 is

$$\boxed{C_3 = \sqrt{\frac{1}{\frac{1}{\alpha}\sin^2\left(\frac{ka}{2}\right)+\frac{a}{2}-\frac{1}{2k}\sin(ka)}}}, \tag{9.4.33}$$

where α is given by Eq. 9.4.20) and k by Eq. 9.4.23. This looks a lot like Eq. 9.4.32. We've just exchanged a cosine for a sine in the first term and a plus for a minus in the third term.

- We've now reached a point where we can't go any further without knowing specific measurements of the finite well. For the well in this example, the width is $a = 0.1$ nm and the depth is $V_0 = 250$ eV, both of which affect how many eigenstates can fit into the well. **Finite wells have a finite number of eigenstates.**

- Using units to match the givens (nm and eV), $\hbar = 6.582 \times 10^{-16}$ eVs and

$$m = 511 \times 10^3 \ \tfrac{\text{eV}}{c^2} = 5.686 \times 10^{-30} \ \tfrac{\text{eV s}^2}{\text{nm}^2}$$

where c is the speed of light (Eq. 5.5.4). That means α or k are

$$\alpha = \tfrac{\sqrt{-2mE}}{\hbar} = \tfrac{5.123}{\text{nm}} \sqrt{\tfrac{-E}{\text{eV}}} \tag{9.4.34}$$

and

$$k = \tfrac{\sqrt{2m(E+V_0)}}{\hbar} = \tfrac{5.123}{\text{nm}} \sqrt{\tfrac{E+250 \text{ eV}}{\text{eV}}} \tag{9.4.35}$$

for any E in electron volts.

- We can't determine specific values for α or k without knowing the possible values of E. This will involve all the boundary conditions at $x = a/2$. For the even solutions,

$$\left\{ \begin{array}{rcccl} \Psi_{\text{even}}\big|_{x=\frac{a}{2}} & = & C_1 e^{-\alpha a/2} & = & C_4 \cos\left(k\tfrac{a}{2}\right) \\ \left.\dfrac{\partial \Psi_{\text{even}}}{\partial x}\right|_{x=\frac{a}{2}} & = & -\alpha C_1 e^{-\alpha a/2} & = & -k C_4 \sin\left(k\tfrac{a}{2}\right) \end{array} \right\}$$

and, if we divide the second equation by the first and move some things around, we get

$$-\alpha = -k \tan\left(k\frac{a}{2}\right)$$

$$\frac{\alpha}{k} = \tan\left(k\frac{a}{2}\right).$$

We know $a = 0.1$ nm as well as α (Eq. 9.4.34) and k (Eq. 9.4.35), so

$$\frac{\tfrac{5.123}{\text{nm}} \sqrt{\tfrac{-E}{\text{eV}}}}{\tfrac{5.123}{\text{nm}} \sqrt{\tfrac{E+250 \text{ eV}}{\text{eV}}}} = \tan\left(\tfrac{5.123}{\text{nm}} \sqrt{\tfrac{E+250 \text{ eV}}{\text{eV}}} \tfrac{0.1 \text{ nm}}{2} \right)$$

$$\sqrt{\frac{-E}{E+250 \text{ eV}}} = \tan\left(0.2562 \sqrt{\frac{E+250 \text{ eV}}{\text{eV}}} \right), \tag{9.4.36}$$

which is the **energy condition** for *even* solutions. Only values of E that satisfy Eq. 9.4.36 are allowed.

- For the odd solutions,

$$\left\{ \begin{array}{rcccl} \Psi_{\text{odd}}\big|_{x=\frac{a}{2}} & = & -C_1 e^{-\alpha a/2} & = & C_4 \sin\left(k\frac{a}{2}\right) \\ \left.\dfrac{\partial \Psi_{\text{odd}}}{\partial x}\right|_{x=\frac{a}{2}} & = & \alpha C_1 e^{-\alpha a/2} & = & kC_4 \cos\left(k\frac{a}{2}\right) \end{array} \right\}$$

and, if we divide the first equation by the second and move some things around, we get

$$-\frac{1}{\alpha} = \frac{1}{k}\tan\left(k\frac{a}{2}\right)$$

$$-\frac{k}{\alpha} = \tan\left(k\frac{a}{2}\right).$$

We know $a = 0.1$ nm as well as α (Eq. 9.4.34) and k (Eq. 9.4.35), so

$$-\frac{\frac{5.123}{\text{nm}}\sqrt{\frac{E+250\text{ eV}}{\text{eV}}}}{\frac{5.123}{\text{nm}}\sqrt{\frac{-E}{\text{eV}}}} = \tan\left(\frac{5.123}{\text{nm}}\sqrt{\frac{E+250\text{ eV}}{\text{eV}}}\frac{0.1\text{ nm}}{2}\right)$$

$$-\sqrt{\frac{E+250\text{ eV}}{-E}} = \tan\left(0.2562\sqrt{\frac{E+250\text{ eV}}{\text{eV}}}\right), \qquad (9.4.37)$$

which is the **energy condition** for *odd* solutions. Only values of E that satisfy Eq. 9.4.37 are allowed.

- Eq. 9.4.36 and 9.4.37 are both transcendental equations (i.e. they "transcend" algebra). This means they are not solvable using algebra, so we're forced to use numerical methods. In Figure 9.14, intersections represent allowed energies and we can see there are only three:

$$\begin{array}{rcl} E_0 & = & -226.0\text{ eV} \\ E_1 & = & -156.1\text{ eV} \\ E_2 & = & -51.28\text{ eV}. \end{array} \qquad (9.4.38)$$

The subscripts are arbitrary, but I've forced "1" to represent the odd solution so it matches everyone's concept of "odd." If the well is deeper (i.e. V_0 is larger), then there are more possible energies. If the well is shallower (i.e. V_0 is smaller), there are fewer possible energies. However, there is *always* an E_0 because α/k will always intersect $\tan(ka/2)$ at least once between zero and $-V_0$.

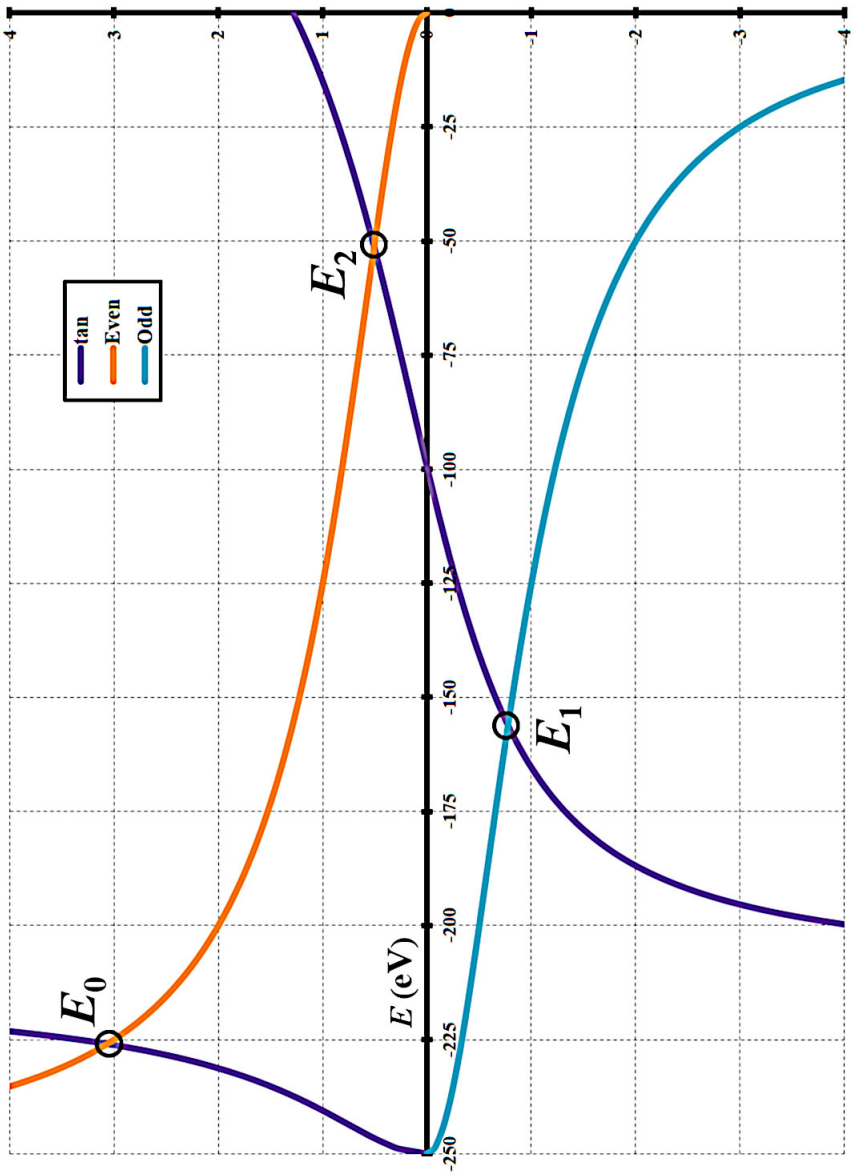


Figure 9.14: This graph shows three functions: $\tan(ka/2)$ ("tan"), α/k ("even"), and $-k/\alpha$ ("odd"). The values shown are for the finite square well given in Example 9.4.3. An intersection of α/k with $\tan(ka/2)$ represents the energy levels of the even solutions. An intersection of $-k/\alpha$ with $\tan(ka/2)$ represents the energy levels of the odd solutions.

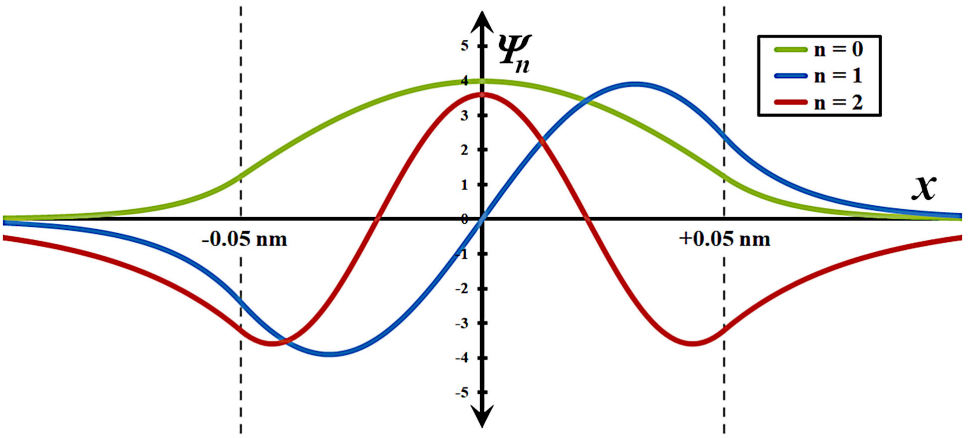


Figure 9.15: These are the *only* three eigenstates in the finite square well given in Example 9.4.3. The vertical dashed lines represent the boundaries of the well.

- With these energies, we can find the three possible eigenstates for this finite well. After finding each α, k, C_1, C_3, and C_4; they are

$$\Psi_0 = \begin{cases} \frac{58.11}{\sqrt{\text{nm}}} e^{77.01x/\text{nm}} & , \text{ if } x < -0.05 \text{ nm} \\ \frac{3.984}{\sqrt{\text{nm}}} \cos\left(\frac{25.11}{\text{nm}} x\right) & , \text{ if } |x| \leq 0.05 \text{ nm} \\ \frac{58.11}{\sqrt{\text{nm}}} e^{-77.01x/\text{nm}} & , \text{ if } x > +0.05 \text{ nm} \end{cases} , \tag{9.4.39}$$

$$\Psi_1 = \begin{cases} -\frac{58.75}{\sqrt{\text{nm}}} e^{64.01x/\text{nm}} & , \text{ if } x < -0.05 \text{ nm} \\ \frac{3.903}{\sqrt{\text{nm}}} \sin\left(\frac{49.63}{\text{nm}} x\right) & , \text{ if } |x| \leq 0.05 \text{ nm} \\ \frac{58.75}{\sqrt{\text{nm}}} e^{-64.01x/\text{nm}} & , \text{ if } x > +0.05 \text{ nm} \end{cases} , \tag{9.4.40}$$

and

$$\Psi_2 = \begin{cases} -\frac{20.09}{\sqrt{\text{nm}}} e^{36.69x/\text{nm}} & , \text{ if } x < -0.05 \text{ nm} \\ \frac{3.597}{\sqrt{\text{nm}}} \cos\left(\frac{72.22}{\text{nm}} x\right) & , \text{ if } |x| \leq 0.05 \text{ nm} \\ -\frac{20.09}{\sqrt{\text{nm}}} e^{-36.69x/\text{nm}} & , \text{ if } x > +0.05 \text{ nm} \end{cases} . \tag{9.4.41}$$

They are all shown in Figure 9.15.

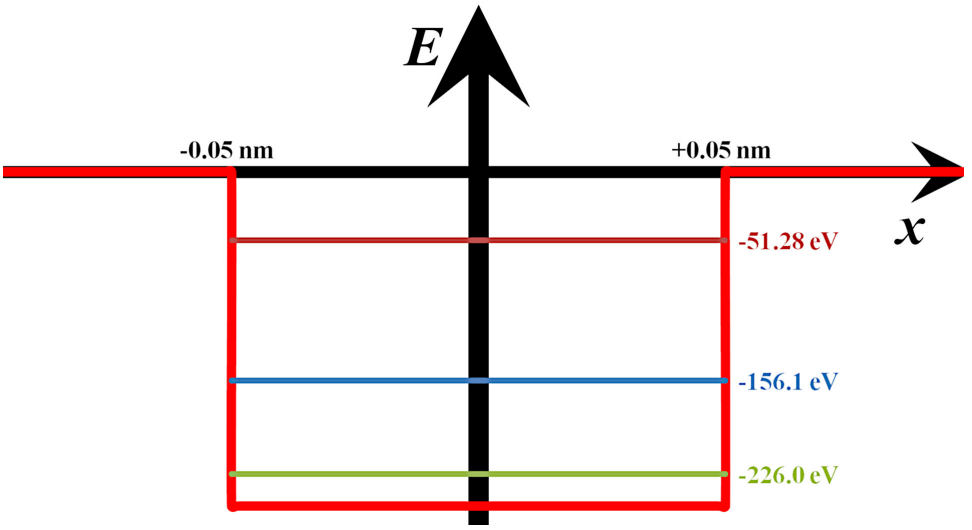


Figure 9.16: This is an **energy level diagram** showing *only* three energy states in the finite square well given in Example 9.4.3. The colors match those used in Figure 9.15.

- The time-evolution factors (Eq. 9.3.12) carry a factor of $E/\hbar$, so

$$U_0(t) = e^{-i(-226.0\text{ eV})t/\left(6.582\times10^{-16}\text{ eVs}\right)} = e^{i\,226.0t/(0.6582\text{ eV fs})} = e^{i\,343.3t/\text{fs}}$$
$$U_1(t) = e^{-i(-156.1\text{ eV})t/\left(6.582\times10^{-16}\text{ eVs}\right)} = e^{i\,156.1t/(0.6582\text{ eV fs})} = e^{i\,237.2t/\text{fs}}$$
$$U_2(t) = e^{-i(-51.28\text{ eV})t/\left(6.582\times10^{-16}\text{ eVs}\right)} = e^{i\,51.28t/(0.6582\text{ eV fs})} = e^{i\,77.91t/\text{fs}},$$

where we've converted to femtoseconds for an aesthetically pleasing exponent. Since the stationary states are given by Eq. 9.3.13, we get

$$\psi_0 = \begin{cases} \frac{58.11}{\sqrt{\text{nm}}} e^{77.01x/\text{nm}} e^{i\,343.3t/\text{fs}} & , \text{ if } x < -0.05\text{ nm} \\ \frac{3.984}{\sqrt{\text{nm}}} \cos\left(\frac{25.11}{\text{nm}}x\right) e^{i\,343.3t/\text{fs}} & , \text{ if } |x| \le 0.05\text{ nm} \\ \frac{58.11}{\sqrt{\text{nm}}} e^{-77.01x/\text{nm}} e^{i\,343.3t/\text{fs}} & , \text{ if } x > +0.05\text{ nm} \end{cases}, \tag{9.4.42}$$

$$\psi_1 = \begin{cases} -\frac{58.75}{\sqrt{\text{nm}}} e^{64.01x/\text{nm}} e^{i\,237.2t/\text{fs}} & , \text{ if } x < -0.05\text{ nm} \\ \frac{3.903}{\sqrt{\text{nm}}} \sin\left(\frac{49.63}{\text{nm}}x\right) e^{i\,237.2t/\text{fs}} & , \text{ if } |x| \le 0.05\text{ nm} \\ \frac{58.75}{\sqrt{\text{nm}}} e^{-64.01x/\text{nm}} e^{i\,237.2t/\text{fs}} & , \text{ if } x > +0.05\text{ nm} \end{cases}, \tag{9.4.43}$$

and

$$\psi_2 = \begin{cases} -\frac{20.09}{\sqrt{\text{nm}}} e^{36.69x/\text{nm}} e^{i\,77.91t/\text{fs}} & , \text{ if } x < -0.05 \text{ nm} \\ \frac{3.597}{\sqrt{\text{nm}}} \cos\left(\frac{72.22}{\text{nm}} x\right) e^{i\,77.91t/\text{fs}} & , \text{ if } |x| \leq 0.05 \text{ nm} \\ -\frac{20.09}{\sqrt{\text{nm}}} e^{-36.69x/\text{nm}} e^{i\,77.91t/\text{fs}} & , \text{ if } x > +0.05 \text{ nm} \end{cases} . \quad (9.4.44)$$

The time-evolution factors have been included on the outside because they are common to all regions.

Example 9.4.4

In Example 9.4.3, an electron was in a finite square well with dimensions $V_0 = 250$ eV and $a = 0.1$ nm. Suppose that electron is in the stationary state ψ_1. What is the probability of finding that electron *inside* the well (i.e. $|x| \leq 0.05$ nm)? What is the probability of finding that electron *outside* the well (i.e. $|x| > 0.05$ nm)?

- The full stationary state is unnecessary since we'll be taking a complex square and

$$\psi^* \psi \, dx = \left(\Psi^* e^{iEt/\hbar}\right)\left(\Psi e^{-iEt/\hbar}\right) = \Psi^* \Psi .$$

All we need to know is the eigenstate Ψ_1, which is given by Eq. 9.4.40. Since we're only interested in the probability inside the well, the state is

$$\Psi_1 = C_3 \sin(kx) = \frac{3.903}{\sqrt{\text{nm}}} \sin\left(\frac{49.63}{\text{nm}} x\right) ,$$

where we've included the symbols for integration generality.

- The probability of finding any particle in a certain region is given by Eq. 9.2.16 in three dimensions. In one dimension, that reduces to

$$P = \int_{x_1}^{x_2} \Psi^* \Psi \, dx \quad (9.4.45)$$

in the region between x_1 and x_2. Inside the well, the boundaries are $x_1 = -a/2$ and $x_2 = +a/2$, so

$$P = \int_{-a/2}^{+a/2} \left[C_3 \sin(kx)\right]^2 \, dx = C_3^2 \int_{-a/2}^{+a/2} \sin^2(kx) \, dx,$$

where we've replaced the complex square with a real square *only* because Ψ is entirely real. Since we've already solved an integral like this in Eq. 9.4.6,

$$P = C_3^2 \left[\frac{x}{2} - \frac{1}{4k}\sin^2(2kx)\right]_{-a/2}^{+a/2} = C_3^2 \left[2\left(\frac{a}{4} - \frac{1}{4k}\sin^2(ka)\right)\right]$$

$$\boxed{P = C_3^2 \left[\frac{a}{2} - \frac{1}{2k}\sin^2(ka)\right]}, \tag{9.4.46}$$

which is true for any odd state in *any* finite square well. Putting the numbers back in, we get

$$P = \left(\frac{3.903}{\sqrt{\text{nm}}}\right)^2 \left[\frac{0.1 \text{ nm}}{2} - \frac{1}{2\left(\frac{49.63}{\text{nm}}\right)}\sin^2\left(\frac{49.63}{\text{nm}}(0.1 \text{ nm})\right)\right],$$

which comes to 0.9105 or 91.05%.

- We could go through another integral to find the probability outside the well, but there's an easier way. Since the probability of finding the electron *somewhere* is 100%, then $P = 100\% - 91.05\% = 8.947\%$ (or 4.474% per side due to symmetry).

Example 9.4.5

In Example 9.4.3, an electron was in a finite square well with dimensions $V_0 = 250$ eV and $a = 0.1$ nm. Suppose that electron is in the stationary state ψ_1. Find $\langle x \rangle$, $\langle x^2 \rangle$, $\langle \mathcal{H} \rangle$, $\langle p \rangle$, and $\langle p^2 \rangle$. Is the Heisenberg Uncertainty Principle (Eq. 9.3.39) satisfied?

- We have two notations which allow us to calculate expectation values: bra-ket notation given by Eq. 9.3.3 and integral notation given by Eq. 9.3.1. If we wanted to find the expectation value of x in bra-ket notation, it would be

$$\langle x \rangle = \langle 1 | \, x \, | 1 \rangle \, ,$$

where $|1\rangle$ is shorthand for $|\psi_1\rangle$. Unfortunately, we have no idea how x operates on $|1\rangle$. We could write $|1\rangle$ in terms of $|x\rangle$ using Eq. 9.3.7, which would give us

$$|1\rangle = \int_{-\infty}^{\infty} |x\rangle \, \langle x | \, |1\rangle \; dx,$$

where we use an integral rather than a sum because x is *continuous*. Bra-ket is far more useful when you're working with *discrete* variables like **spin**. Since continuous variables force us into the realm of integrals, we might as well just use integral notation.

- In integral notation, the expectation value of some observable Q is given by

$$\langle Q \rangle = \int_{-\infty}^{+\infty} \psi^* Q \psi \, dx. \tag{9.4.47}$$

However, if Q is *not* dependent on time, the full stationary state is unnecessary since

$$\psi^* Q \psi = \left(\Psi^* e^{iEt/\hbar} \right) \left(Q \Psi e^{-iEt/\hbar} \right) = \Psi^* Q \Psi.$$

The eigenstate is also entirely real, so $\Psi^* = \Psi$ and

$$\langle Q \rangle = \int_{-\infty}^{+\infty} \Psi Q \Psi \, dx.$$

Since Ψ is different over different values of x,

$$\langle Q \rangle = \int_{-\infty}^{-a/2} \Psi Q \Psi \, dx + \int_{-a/2}^{+a/2} \Psi Q \Psi \, dx + \int_{+a/2}^{+\infty} \Psi Q \Psi \, dx \tag{9.4.48}$$

where Q is still an arbitrary observable. This will work for all expectation values in this example.

- Using Eq. 9.4.48 for $\langle x \rangle$ gives us

$$\begin{aligned} \langle x \rangle &= \int_{-\infty}^{-a/2} [C_1 e^{\alpha x}]\, x\, [C_1 e^{\alpha x}]\ dx \\ &+ \int_{-a/2}^{+a/2} [C_3 \sin(kx)]\, x\, [C_3 \sin(kx)]\ dx \\ &+ \int_{+a/2}^{+\infty} [-C_1 e^{-\alpha x}]\, x\, [-C_1 e^{-\alpha x}]\ dx \end{aligned}$$

  where $a = 0.1$ nm all other constants are given in Eq. 9.4.40. If we keep the constants general, then our result will apply to any odd state in *any* finite square well. Since an operation of x does not change Ψ, we get

$$\langle x \rangle = C_1^2 \int_{-\infty}^{-a/2} x e^{2\alpha x}\, dx + C_3^2 \int_{-a/2}^{+a/2} x \sin^2(kx)\ dx + C_1^2 \int_{+a/2}^{+\infty} x e^{-2\alpha x}\, dx$$

  and we have three integrals we can solve analytically.

- Luckily, we can avoid solving them using a couple shortcuts. The middle integrand, $x \sin^2(kx)$, is an odd function (i.e. $f_{\text{odd}}(-x) = -f_{\text{odd}}(x)$). Odd functions have the property

$$\int_{-b}^{+b} f_{\text{odd}}(x)\ dx = 0 \tag{9.4.49}$$

  for any b and any f_{odd}, so the second integral is also zero. This leaves us with only

$$\langle x \rangle = C_1^2 \int_{-\infty}^{-a/2} x e^{2\alpha x}\, dx + C_1^2 \int_{+a/2}^{+\infty} x e^{-2\alpha x}\, dx$$

  to solve. If, in the first term, we reverse the limits of integration and say $x = -x'$, then

$$\begin{aligned} \int_{-\infty}^{-a/2} x e^{2\alpha x}\, dx &= -\int_{-a/2}^{-\infty} x e^{2\alpha x}\, dx \\ &= -\int_{+a/2}^{+\infty} (-x')\, e^{2\alpha(-x')}\, d(-x') \\ &= -\int_{+a/2}^{+\infty} x' e^{-2\alpha x'}\, dx'. \end{aligned}$$

Since x' is just a label, it could just as easily be x and

$$\int_{-\infty}^{-a/2} xe^{2\alpha x}\,dx = -\int_{+a/2}^{+\infty} xe^{-2\alpha x}\,dx.$$

This means the two remaining terms cancel each other and $\boxed{\langle x\rangle = 0}$, which is actually true for all the states in any finite square well. Remember, an expectation value is just a *weighted average* of all possible values. Due to the symmetry of the eigenstate, the electron is just as likely to be found at a negative value for x as it is a positive value for x.

- The same tricks wont work on $\langle x^2\rangle$. Using Eq. 9.4.48 gives us

$$\begin{aligned}\langle x^2\rangle \;=\;& \int_{-\infty}^{-a/2} [C_1 e^{\alpha x}]\, x^2\, [C_1 e^{\alpha x}]\;dx \\ &+ \int_{-a/2}^{+a/2} [C_3\sin(kx)]\, x^2\, [C_3\sin(kx)]\;dx \\ &+ \int_{+a/2}^{+\infty} [-C_1 e^{-\alpha x}]\, x^2\, [-C_1 e^{-\alpha x}]\;dx\end{aligned}$$

where $a = 0.1$ nm all other constants are given in Eq. 9.4.40. If we keep the constants general, then our result will apply to any odd state in *any* finite square well. Since an operation of x^2 does not change Ψ, we get

$$\langle x^2\rangle = C_1^2\int_{-\infty}^{-a/2} x^2 e^{2\alpha x}\,dx + C_3^2\int_{-a/2}^{+a/2} x^2\sin^2(kx)\;dx + C_1^2\int_{+a/2}^{+\infty} x^2 e^{-2\alpha x}\,dx$$

and we have three integrals we can solve analytically.

- Unfortunately, we have to solve these integrals, but we may still be able to save a little time. The middle integrand, $x^2\sin^2(kx)$, is an even function (i.e. $f_{\text{even}}(-x) = f_{\text{even}}(x)$). Even functions have the property

$$\int_{-b}^{+b} f_{\text{even}}(x)\;dx = 2\int_{0}^{+b} f_{\text{even}}(x)\;dx \tag{9.4.50}$$

for any b and any f_{even}. Even and odd functions show up a lot in quantum mechanics, so you should get used to using Eqs. 9.4.49 and 9.4.50.

Additionally if, in the first term, we reverse the limits of integration and say $x = -x'$, then

$$\begin{aligned}\int_{-\infty}^{-a/2} x^2 e^{2\alpha x}\,dx &= -\int_{-a/2}^{-\infty} x^2 e^{2\alpha x}\,dx \\ &= -\int_{+a/2}^{+\infty} (-x')^2\, e^{2\alpha(-x')}\,d(-x') \\ &= +\int_{+a/2}^{+\infty} (x')^2\, e^{-2\alpha x'}\,dx'.\end{aligned}$$

Since x' is just a label, it could just as easily be x and

$$\int_{-\infty}^{-a/2} x e^{2\alpha x}\,dx = \int_{+a/2}^{+\infty} x^2 e^{-2\alpha x}\,dx.$$

This means the first and last terms are equal, so the expectation value simplifies to

$$\langle x^2 \rangle = 2C_3^2 \int_0^{a/2} x^2 \sin^2(kx)\;dx + 2C_1^2 \int_{a/2}^{\infty} x^2 e^{-2\alpha x}\,dx,$$

where we've completely eliminated negative values of x.

- We could look up these integrals in an integral table, but where's the fun that? It certainly wouldn't fit the pattern in this book. My favorite method for integrals like this is called the **tabular integration method**, which is a way to do integration by parts multiple times all at once. You need to form columns, taking derivatives in one and integrals in the other (alternating signs) until one of them goes to zero. Using this method and Eq. 9.4.6, the first integral is

$$\int_0^{a/2} x^2 \sin^2(kx)\;dx = \left\{ \begin{array}{lll} +x^2 & \times + & \left(\frac{x}{2} - \frac{\sin(2kx)}{4k}\right) \\ +2x & \times - & \left(\frac{x^2}{4} + \frac{\cos(2kx)}{8k^2}\right) \\ +2 & \times + & \left(\frac{x^3}{12} + \frac{\sin(2kx)}{16k^3}\right) \end{array} \right|_0^{a/2},$$

where $u = x^2$ and $dv = \sin^2(kx)\;dx$ in the common notation for integration by parts. Simplifying, we get

$$\int_0^{a/2} x^2 \sin^2(kx)\;dx = -\frac{x^2 \sin(2kx)}{4k} - \frac{x\cos(2kx)}{4k^2} + \frac{x^3}{6} + \frac{\sin(2kx)}{8k^3}\bigg|_0^{a/2}$$

$$\int_0^{a/2} x^2 \sin^2(kx)\, dx = -\tfrac{a^2\sin(ka)}{16k} - \tfrac{a\cos(ka)}{8k^2} + \tfrac{a^3}{48} + \tfrac{\sin(ka)}{8k^3}. \quad (9.4.51)$$

Using the tabular method and Eq. 9.4.30, the second integral is

$$\int_{a/2}^{\infty} x^2 e^{-2\alpha x}\, dx = \left\{ \begin{array}{lll} +x^2 & \times + & \left(-\tfrac{e^{-2\alpha x}}{2\alpha}\right) \\ +2x & \times - & \left(\tfrac{e^{-2\alpha x}}{4\alpha^2}\right) \\ +2 & \times + & \left(-\tfrac{e^{-2\alpha x}}{8\alpha^3}\right) \end{array} \right|_{a/2}^{\infty},$$

where $u = x^2$ and $dv = e^{-2\alpha x}dx$ in the common notation for integration by parts. Simplifying, we get

$$\int_{a/2}^{\infty} x^2 e^{-2\alpha x}\, dx = -\left(\frac{x^2}{2\alpha} + \frac{x}{2\alpha^2} + \frac{1}{4\alpha^3}\right) e^{-2\alpha x}\Big|_{a/2}^{\infty}$$

$$\int_{a/2}^{\infty} x^2 e^{-2\alpha x}\, dx\, dx = \left(\frac{a^2}{8\alpha} + \frac{a}{4\alpha^2} + \frac{1}{4\alpha^3}\right) e^{-\alpha a}. \quad (9.4.52)$$

- If we substitute Eqs. 9.4.51 and 9.4.52 back into the expectation value, then

$$\begin{aligned} \langle x^2 \rangle &= 2C_3^2 \left[-\tfrac{a^2\sin(ka)}{16k} - \tfrac{a\cos(ka)}{8k^2} + \tfrac{a^3}{48} + \tfrac{\sin(ka)}{8k^3} \right] \\ &+ 2C_1^2 \left[\left(\tfrac{a^2}{8\alpha} + \tfrac{a}{4\alpha^2} + \tfrac{1}{4\alpha^3} \right) e^{-\alpha a} \right] \end{aligned}$$

$$\boxed{\begin{aligned} \langle x^2 \rangle &= C_3^2 \left[-\tfrac{a^2\sin(ka)}{8k} - \tfrac{a\cos(ka)}{4k^2} + \tfrac{a^3}{24} + \tfrac{\sin(ka)}{4k^3} \right] \\ &+ C_1^2 \left[\left(\tfrac{a^2}{4\alpha} + \tfrac{a}{2\alpha^2} + \tfrac{1}{2\alpha^3} \right) e^{-\alpha a} \right], \end{aligned}} \quad (9.4.53)$$

which applies to any odd state in *any* finite well. If we put all the numbers in from Eq. 9.4.40, then $\langle x^2 \rangle = 4.990 \times 10^{-4}$ nm^2.

- Since Ψ_1 is an eigenstate of $\mathcal{H}$, we know the energy is definite (and *discrete*), which makes our work much easier. Mathematically, we say

$$\mathcal{H}\,|1\rangle = E_1\,|1\rangle\,,$$

so

$$\langle \mathcal{H} \rangle = \langle 1| \mathcal{H} |1\rangle = \langle 1| E_1 |1\rangle = E_1 \langle 1| |1\rangle = E_1 = -156.1 \text{ eV}.$$

We know the value of this from Eq. 9.4.38. This method works for all eigenstates,

$$\langle \mathcal{H} \rangle = \langle n| \mathcal{H} |n\rangle = E_n \,, \tag{9.4.54}$$

where n is the state number.

- We should expect $\langle p \rangle = 0$ since the electron would just as likely be traveling right as it would left, but we'll write out the integrals just to be safe. Using Eq. 9.4.48 gives us

$$\begin{aligned} \langle p \rangle \;\; = \;\; & \int_{-\infty}^{-a/2} [C_1 e^{\alpha x}]\, p\, [C_1 e^{\alpha x}] \; dx \\ + \;\; & \int_{-a/2}^{+a/2} [C_3 \sin(kx)]\, p\, [C_3 \sin(kx)] \; dx \\ + \;\; & \int_{+a/2}^{+\infty} [-C_1 e^{-\alpha x}]\, p\, [-C_1 e^{-\alpha x}] \; dx \end{aligned}$$

where $a = 0.1$ nm all other constants are given in Eq. 9.4.40. If we keep the constants general, then our result will apply to any odd state in *any* finite square well. Since p (Eq. 9.2.3) is a derivative,

$$p = -i\hbar \frac{\partial}{\partial x} \tag{9.4.55}$$

in one dimension, its operation does change Ψ. Pulling out all constants, we get

$$\begin{aligned} \langle p \rangle \;\; = \;\; & -i\hbar \Big[C_1^2 \int_{-\infty}^{-a/2} e^{\alpha x} \tfrac{\partial}{\partial x} e^{\alpha x} \; dx \\ & + C_3^2 \int_{-a/2}^{+a/2} \sin(kx)\, \tfrac{\partial}{\partial x} \sin(kx) \; dx \\ & + C_1^2 \int_{+a/2}^{+\infty} e^{-\alpha x} \tfrac{\partial}{\partial x} e^{-\alpha x} \; dx \Big] \end{aligned}$$

and evaluating each of the derivatives gives

$$\begin{aligned} \langle p \rangle \;\; = \;\; & -i\hbar \Big[\alpha C_1^2 \int_{-\infty}^{-a/2} e^{2\alpha x} \; dx \\ & + k C_3^2 \int_{-a/2}^{+a/2} \sin(kx) \cos(kx) \; dx \\ & - \alpha C_1^2 \int_{+a/2}^{+\infty} e^{-2\alpha x} \; dx \Big] \,. \end{aligned}$$

We now have three integrals we can solve analytically.

- Luckily, we can avoid solving them using a couple shortcuts. The middle integrand, $\sin(kx)\cos(kx)$, is an odd function (i.e. $f_{\text{odd}}(-x) = -f_{\text{odd}}(x)$). Eq. 9.4.49 tells us the second integral is zero. This leaves us with only

$$\langle p \rangle = -i\hbar \left[\alpha C_1^2 \int_{-\infty}^{-a/2} e^{2\alpha x}\, dx - \alpha C_1^2 \int_{+a/2}^{+\infty} e^{-2\alpha x}\, dx \right]$$

$$\langle p \rangle = -i\hbar\alpha \left[C_1^2 \int_{-\infty}^{-a/2} e^{2\alpha x}\, dx - C_1^2 \int_{+a/2}^{+\infty} e^{-2\alpha x}\, dx \right]$$

to solve. The remaining integrals are just the probability of finding the electron outside the finite well on either side. We know from Example 9.4.4 these two integrals have the same result (0.04474), so they cancel each other and $\boxed{\langle p \rangle = 0}$, which is actually true for *all* the states in *any* finite square well.

- The only expectation value left is $\langle p^2 \rangle$. We could go through the integrals, but there's an easier way. Eqs. 9.2.2 and 9.2.4 tell us that

$$\mathcal{H} = \frac{p^2}{2m} + V = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x) \tag{9.4.56}$$

in one dimension for non-relativistic particles. If we take the expectation value, then

$$\langle \mathcal{H} \rangle = \left\langle \frac{p^2}{2m} + V \right\rangle = \frac{\langle p^2 \rangle}{2m} + \langle V \rangle$$

since the expectation operator is linear. We know $V = -V_0$ is constant and we already found in Eq. 9.4.54 that $\langle \mathcal{H} \rangle = E_n$ is definite. Rearranging, we get

$$E_n = \frac{\langle p^2 \rangle}{2m} - V_0$$

$$\Rightarrow \boxed{\langle p^2 \rangle = 2m\,(E_n + V_0) = (\hbar k)^2}, \tag{9.4.57}$$

where the appearance of $\hbar k$ should be no surprise. Its value in this example is $\langle p^2 \rangle = 2463\left(\frac{\hbar}{\text{nm}}\right)^2$, where the value of k comes from Eq. 9.4.40.

- It should also be no surprise that the value of p^2 is *definite* because, in this example, $\mathcal{H}$ and p^2 commute. In general, the commutator is

$$\left[\mathcal{H}, p^2\right] = \hbar^2 \frac{\partial^2 V}{\partial x^2},$$

but V is constant so $[\mathcal{H}, p^2] = 0$. This means they are compatible and share a set of eigenstates (i.e. it is possible to find the particle in a stationary state of both observables). According to the general uncertainty principle (Eq. 9.3.31), we can predict them both with precision.

- We can also confirm the more specific uncertainty principle for position and momentum (Eq. 9.3.39) with a little computation. The standard deviations, according to Eq. 9.3.23, are

$$\sigma_x = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \sqrt{4.990 \times 10^{-4} \text{ nm}^2 - 0} = 0.02234 \text{ nm}$$

and

$$\sigma_p = \sqrt{\langle p^2 \rangle - \langle p \rangle^2} = \sqrt{2463 \left(\tfrac{\hbar}{\text{nm}}\right)^2 - 0} = 49.63 \, \tfrac{\hbar}{\text{nm}}.$$

- Remember, this a measure of variation from the expectation value. For example, if we were to do 1000 *identical* experiments, then we can expect two things:

  1. the *weighted* average of all the x-values will be roughly $\langle x \rangle$ and
  2. roughly 68% of x-values will be between $\langle x \rangle - \sigma_x$ and $\langle x \rangle + \sigma_x$,

  where "68%" is a purely statistical result for normally distributed systems (i.e. *most* quantum systems). For this finite well, that 68% would be found between $x = -0.02234$ nm and $x = +0.02234$ nm. This is all we can really predict about x.

- Multiplying these two standard deviations, gives us

$$\sigma_x \sigma_p = (0.02234 \text{ nm}) \left(49.63 \, \tfrac{\hbar}{\text{nm}}\right) = 1.109\hbar,$$

which is greater than $\hbar/2$. This is consistent with the uncertainty principle (Eq. 9.3.39).

Table 9.1: This is a summary of all the calculated values in Example 9.4.5.

Quantity		Value
$\langle x \rangle$	=	0
$\langle x^2 \rangle$	=	4.990×10^{-4} nm^2
$\langle \mathcal{H} \rangle$	=	-156.1 eV
$\langle p \rangle$	=	0
$\langle p^2 \rangle$	=	$2463 \left(\frac{\hbar}{\text{nm}}\right)^2$
σ_x	=	0.02234 nm
σ_p	=	$49.63 \frac{\hbar}{\text{nm}}$
$\sigma_x \sigma_p$	=	$1.109\hbar$

Harmonic Oscillator

Another unrealistic quality, present in both the infinite and finite square wells, is discontinuity. The potential energy function changes abruptly at the boundaries. If we want a more realistic potential energy function, then we need it to be continuous over space. The upside is we wont have to write the quantum states in piecewise form. The downside is they can be tricky or sometimes even impossible to solve analytically.

The simplest of these continuous models is the **harmonic oscillator**. In one dimension, the potential energy function can be written as

$$\boxed{V(x) = \frac{1}{2}kx^2 = \frac{1}{2}m\omega^2 x^2}, \tag{9.4.58}$$

where k is the classical elastic constant and $\omega = \sqrt{k/m}$ is the angular frequency. Note that we've chosen V to be positive everywhere, but you can shift the function up or down as needed without changing the force experienced by the particle. You can write this in three dimensions as

$$V(x, y, z) = \frac{1}{2}m\omega^2 \left(x^2 + y^2 + z^2\right) \tag{9.4.59}$$

in Cartesian coordinates. It can also be written in spherical coordinates as

$$V(r) = \frac{1}{2}m\omega^2 r^2, \tag{9.4.60}$$

but only if the oscillation is isotropic (i.e. independent of direction). If the oscillations are different in different directions, then we're forced to use Eq. 9.4.59.

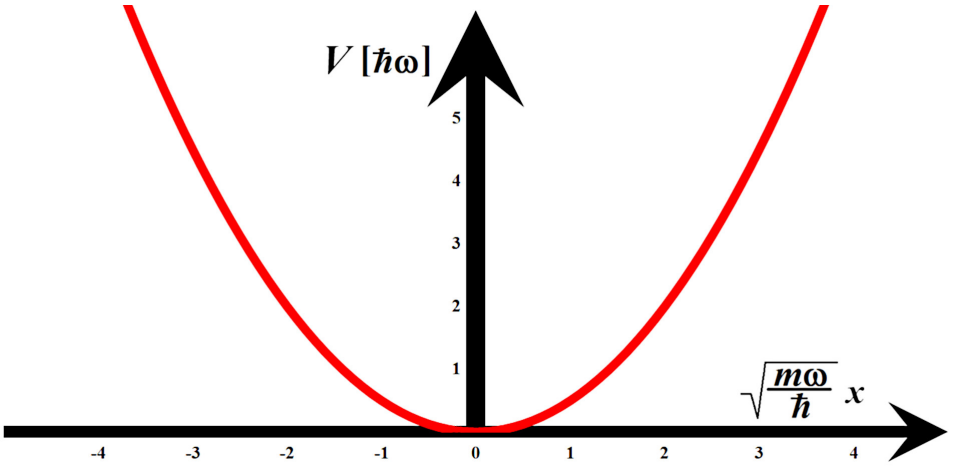


Figure 9.17: This shows the (one-dimensional) harmonic oscillator's potential energy (Eq. 9.4.58) graphed against position. Values on the vertical axis are in units of $\hbar\omega$ and values on the horizontal axis are for $\sqrt{\frac{m\omega}{\hbar}}\,x$ (no unit) rather than x for generality.

Recall from Section 9.1, that Max Planck solved the black body radiation problem by assuming the light-emitting object was made of very small oscillators. The result was that light was emitted in packets called photons, each with an energy of

$$E_{\text{photon}} = nhf = n\hbar\omega, \tag{9.4.61}$$

where n is a whole number (i.e. $n = 1, 2, 3, \ldots$). Planck made this discovery *before* we were even sure atoms existed, which is impressive. These days, we happen to know that atoms emit light when electrons transition between energy levels, so we should expect the energy levels of the harmonic oscillator to differ by Eq. 9.4.61.

Example 9.4.6

What are the stationary states (and corresponding energies) for a non-relativistic particle behaving as a harmonic oscillator?

- The time-independent Schrödinger equation (Eq. 9.3.14) will take the form

$$\mathcal{H}\Psi = -\frac{\hbar^2}{2m}\vec{\nabla}^2\Psi + V\Psi = E\Psi$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}+\frac{1}{2}m\omega^2x^2\Psi=E\Psi, \tag{9.4.62}$$

where $\vec{\nabla}^2$ is only in one dimension, x. There are a few ways to solve a differential equation like this. Many formal texts on the topic use something called **ladder operators** because it's considered "elegant." However, I find it to be more convoluted and inefficient than elegant. I prefer the more direct method of using a **power series solution**.

- Any smooth continuous function (i.e. any state function) can be written as an infinite power series, so it is guaranteed

$$\Psi(x)=\sum_{j=0}^{\infty}a_jx^j \tag{9.4.63}$$

will be a solution to Eq. 9.4.62. This power series could represent *any* function in its current form since the constant coefficients, a_j, are not specified. We can do a little better though.

- First, we're going to choose better variables. Let's say

$$\chi=\sqrt{\frac{m\omega}{\hbar}}\,x \tag{9.4.64}$$

and

$$\varepsilon=\frac{E}{\hbar\omega} \tag{9.4.65}$$

just like in Figure 9.17. Neither χ nor ε have a unit, which makes them very convenient for anything we might have to do numerically. We can apply the chain rule for derivatives (Eq. 3.1.2),

$$-\frac{\hbar^2}{2m}\left(\frac{\partial\chi}{\partial x}\frac{\partial}{\partial\chi}\right)^2\Psi+\frac{1}{2}m\omega^2x^2\Psi=E\Psi,$$

and substitute from χ and ε to get

$$-\frac{\hbar^2}{2m}\left(\sqrt{\frac{m\omega}{\hbar}}\frac{\partial}{\partial\chi}\right)^2\Psi+\frac{1}{2}m\omega^2\left(\sqrt{\frac{\hbar}{m\omega}}\,\chi\right)^2\Psi=\hbar\omega\varepsilon\Psi$$

$$-\frac{1}{2}\hbar\omega\frac{\partial^2\Psi}{\partial\chi^2}+\frac{1}{2}\hbar\omega\chi^2\Psi=\hbar\omega\varepsilon\Psi$$

$$\frac{\partial^2\Psi}{\partial\chi^2}-\chi^2\Psi=-2\varepsilon\Psi.$$

Moving all non-derivative terms to the right, this becomes

$$\frac{\partial^2\Psi}{\partial\chi^2}=\left(\chi^2-2\varepsilon\right)\Psi, \tag{9.4.66}$$

where the 2ε is what's forcing us to use a power series solution.

- Second, we do know *a little* about what the function looks like. As $\chi\to\infty$,

$$\frac{\partial^2\Psi}{\partial\chi^2}\to\chi^2\Psi, \tag{9.4.67}$$

since 2ε is constant. That means $\Psi(\chi)$ should include a factor of $e^{-\chi^2/2}$, which dominates over everything else for large χ. Therefore,

$$\Psi(\chi)=u(\chi)\ e^{-\chi^2/2}, \tag{9.4.68}$$

where u is the "everything else" and must become insignificant for large χ. Technically, $e^{+\chi^2/2}$ is also a solution to Eq. 9.4.67, but we ignored it since Ψ must be finite.

- Last, we write Eq. 9.4.66 in terms of u rather than Ψ. Substituting in Eq. 9.4.68, we get

$$\begin{aligned}
\frac{\partial}{\partial\chi}\frac{\partial}{\partial\chi}\left(u\,e^{-\chi^2/2}\right) &= \left(\chi^2-2\varepsilon\right)u\,e^{-\chi^2/2}\\
\frac{\partial}{\partial\chi}\left(e^{-\chi^2/2}\frac{\partial u}{\partial\chi}-u\chi\,e^{-\chi^2/2}\right) &= \left(\chi^2-2\varepsilon\right)u\,e^{-\chi^2/2}\\
\left(-\chi\frac{\partial u}{\partial\chi}+\frac{\partial^2u}{\partial\chi^2}-\chi\frac{\partial u}{\partial\chi}-u+u\chi^2\right)e^{-\chi^2/2} &= \left(\chi^2-2\varepsilon\right)u\,e^{-\chi^2/2}.
\end{aligned}$$

If cancel the $e^{-\chi^2/2}$ and group like terms, then

$$-\chi\frac{\partial u}{\partial\chi}+\frac{\partial^2u}{\partial\chi^2}-\chi\frac{\partial u}{\partial\chi}-u+u\chi^2=\chi^2u-2\varepsilon u$$

$$\frac{\partial^2 u}{\partial \chi^2} - 2\chi \frac{\partial u}{\partial \chi} + (2\varepsilon - 1)\, u = 0. \tag{9.4.69}$$

Setting everything equal to zero is convenient for what we're about to do.

- Now we use the power series solution,

$$u(\chi) = \sum_{j=0}^{\infty} a_j \chi^j, \tag{9.4.70}$$

to find u. Judging from Eq. 9.4.69, we're going to need a first and second derivative, so

$$\frac{\partial u}{\partial \chi} = \sum_{j=1}^{\infty} j a_j \chi^{j-1}$$

and

$$\frac{\partial^2 u}{\partial \chi^2} = \sum_{j=2}^{\infty} j\,(j-1)\, a_j \chi^{j-2}.$$

The lowest value of j goes up in each sum because the derivative of a constant (e.g. the first term in the sum) is zero. If we substitute all three of these into Eq. 9.4.69, then

$$\sum_{j=2}^{\infty} j\,(j-1)\, a_j \chi^{j-2} - 2\chi \sum_{j=1}^{\infty} j a_j \chi^{j-1} + (2\varepsilon - 1) \sum_{j=0}^{\infty} a_j \chi^j = 0$$

$$\sum_{j=2}^{\infty} j\,(j-1)\, a_j \chi^{j-2} + \sum_{j=1}^{\infty} -2j a_j \chi^j + \sum_{j=0}^{\infty} (2\varepsilon - 1)\, a_j \chi^j = 0.$$

- Unfortunately, we can't combine the sums properly until all the powers of χ are the same *and* all the lower limits are the same. Since j is just a label, we could easily replace every j in the first sum with $j + 2$,

$$\sum_{j+2=2}^{\infty} (j+2)\,(j+2-1)\, a_{j+2} \chi^{j+2-2} + \sum_{j=1}^{\infty} -2j a_j \chi^j + \sum_{j=0}^{\infty} (2\varepsilon - 1)\, a_j \chi^j = 0$$

$$\sum_{j=0}^{\infty}(j+2)(j+1)\,a_{j+2}\chi^j+\sum_{j=1}^{\infty}-2ja_j\chi^j+\sum_{j=0}^{\infty}(2\varepsilon-1)\,a_j\chi^j=0$$

to make the powers the same. To make the lower limits the same, we could just separate the $j=0$ terms from the first and last sum. However, in this case, the second sum has a factor of j, so adding a $j=0$ to that sum would just be adding a zero term (i.e. *voodoo math*). This gives us

$$\sum_{j=0}^{\infty}(j+2)(j+1)\,a_{j+2}\chi^j+\sum_{j=0}^{\infty}-2ja_j\chi^j+\sum_{j=0}^{\infty}(2\varepsilon-1)\,a_j\chi^j=0$$

$$\sum_{j=0}^{\infty}\left[(j+2)(j+1)\,a_{j+2}-2ja_j+(2\varepsilon-1)\,a_j\right]\chi^j=0.$$

- The only way this sum can *always* be zero is when the coefficients are *all* zero. Therefore,

$$(j+2)(j+1)\,a_{j+2}-2ja_j+(2\varepsilon-1)\,a_j=0$$

$$a_{j+2}=\frac{2j+1-2\varepsilon}{(j+2)(j+1)}\,a_j, \tag{9.4.71}$$

which is called a **recursion formula**. It determines all the coefficients in the sum as long as you know two of them, a_0 for the even values of j (i.e. $a_2, a_4, a_6, \ldots$) and a_1 for the odd values of j (i.e. $a_3, a_5, a_7, \ldots$). If this is going to represent a wave function though, there should only be *one* unknown coefficient: the **normalization constant**.

- Remember Eq. 9.4.18? When a potential energy function is symmetric (i.e. $V(-x)=V(x)$), we know the solutions to the time-independent Schrödinger equation (Eq. 9.3.14) must be *even* or *odd*. Eq. 9.4.68 tells us $\Psi=u\,e^{-\chi^2/2}$ and, since $e^{-\chi^2/2}$ is always even, it is up to u to determine the parity of Ψ. Since a_0 is for *even* terms and a_1 is for *odd* terms, we can conclude that one of them is *always* zero in each solution.

  - If $a_0 \neq 0$, then $a_1 = 0$.
  - If $a_1 \neq 0$, then $a_0 = 0$.

  That means, for any particular solution, you really only have one unknown coefficient: either a_0 or a_1, never both. Which ever one is nonzero is your normalization constant.

- Note: **All eigenstates must be continuous (and finite) over all space and must have a first derivative that is continuous (and finite) over all space.** This is because its second derivative must exist in all space due to Schrödinger equation.

- We have to make sure $\Psi \nrightarrow \infty$ as $\chi \to \infty$. The power series for u has an infinite number of terms, but that alone does not pose a problem. If the coefficients, a_j, get smaller in *just the right way* as j get larger, then the series *can* converge to a finite function. We have to check to make sure. As $j \to \infty$,

$$a_{j+2} \to \frac{2j}{(j)\,(j)}\, a_j = \frac{2}{j}\, a_j,$$

  which comes from Eq. 9.4.71. As $\chi \to \infty$, larger j terms matter more, so we could say

$$\begin{aligned} u_{\text{even}} \quad &\to \quad a_0 + \tfrac{2}{2}a_0\chi^2 + \tfrac{2}{4}\left(\tfrac{2}{2}a_0\right)\chi^4 + \tfrac{2}{6}\left(\tfrac{2}{4}\left(\tfrac{2}{2}a_0\right)\right)\chi^6 + \ldots \\ &= a_0\left(1 + \tfrac{1}{1}\chi^{2*1} + \tfrac{1}{2*1}\chi^{2*2} + \tfrac{1}{3*2*1}\chi^{2*3} + \ldots\right) \\ &= a_0 \sum_{\ell=0}^{\infty} \frac{1}{\ell!}\chi^{2\ell} = a_0 \sum_{\ell=0}^{\infty} \frac{1}{\ell!}\left(\chi^2\right)^{\ell} = a_0 e^{\chi^2} \end{aligned}$$

  and, therefore,

$$\Psi_{\text{even}} \to a_0 e^{\chi^2} e^{-\chi^2/2} = a_0 e^{+\chi^2/2}.$$

- As $\chi \to \infty$, $\Psi_{\text{even}} \to \infty$, so this is big problem. The only possible conclusion is that the power series *doesn't* have an infinite number of terms. In other words, there must be some $j_{\max}$ such that $a_{(j_{\max}+2)} = 0$. Using Eq. 9.4.71, we get

$$0 = \frac{2j_{\max} + 1 - 2\varepsilon}{(j_{\max}+2)\,(j_{\max}+1)}\, a_{j_{\max}}$$

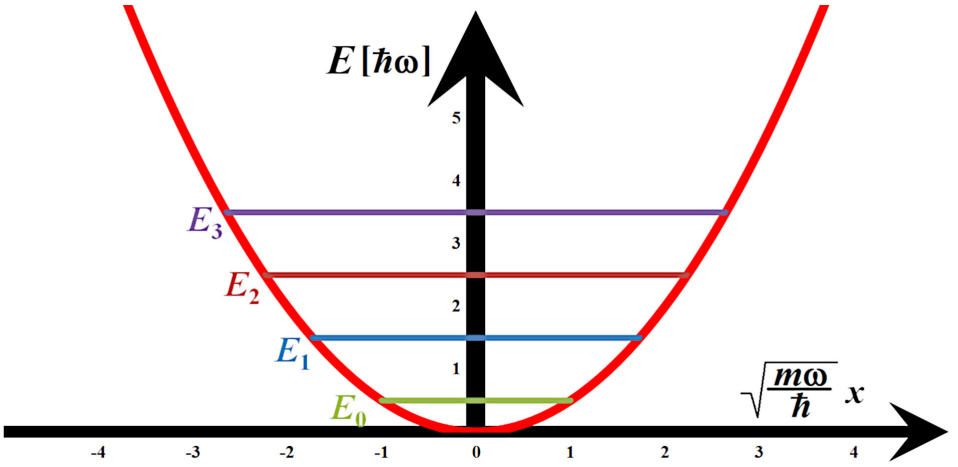


Figure 9.18: This is an **energy level diagram** showing the first four energy states in the (one-dimensional) harmonic oscillator. Values on the vertical axis are in units of $\hbar\omega$ and values on the horizontal axis are for $\sqrt{\frac{m\omega}{\hbar}}\,x$ (no unit) rather than x for generality. The colors match those used in Figure 9.19.

$$0 = 2j_{\max} + 1 - 2\varepsilon \quad \Rightarrow \varepsilon = j_{\max} + \frac{1}{2}.$$

We know ε is related to E by Eq. 9.4.65, so this is just the energy in terms of $j_{\max}$. The convention we've chosen in this chapter is to use n to number energy levels, so

$$\boxed{E_n = \left(n + \frac{1}{2}\right)\hbar\omega} \tag{9.4.72}$$

where $n = j_{\max}$. These energy values all differ by an integer multiple of $\hbar\omega$, which is consistent with Planck's result (Eq. 9.4.61).

- Eq. 9.4.72 allows to write the recursion formula (Eq. 9.4.71) as

$$a_{j+2} = \frac{2j + 1 - (2n + 1)}{(j + 2)(j + 1)}\, a_j$$

$$\boxed{a_{j+2} = \frac{-2(n - j)}{(j + 2)(j + 1)}\, a_j}, \tag{9.4.73}$$

which is in terms of n. Using this one will save us a little time calculating coefficients.

- Now that we have the energy levels and a way to determine coefficients, we should be able to find the eigenstates. Combining Eqs. 9.4.68 and 9.4.70, we get

$$\Psi_n = u_n e^{-\chi/2} = \left[\sum_{j=0}^{n} a_j \chi^j\right] e^{-\chi^2/2},$$

where each a_j is determined by recursion (Eq. 9.4.73). This isn't very efficient though. Ideally, we'd like to write out Ψ_n without having to use recursion, so let's look for a pattern.

- If we expand u_n for *even* solutions, then

$$\begin{aligned} u_{\text{even}} &= \sum_{j=0}^{n} a_j \chi^j = \sum_{\ell=0}^{n/2} a_{2\ell} \chi^{2\ell} \\ &= a_0 + a_2\chi^2 + a_4\chi^4 + a_6\chi^6 + \ldots + a_n\chi^n \\ &= a_0 + \tfrac{-2n}{2*1} a_0\chi^2 + \tfrac{-2(n-2)}{4*3}\tfrac{-2n}{2*1} a_0\chi^4 + \tfrac{-2(n-4)}{6*5}\tfrac{-2(n-2)}{4*3}\tfrac{-2n}{2*1} a_0\chi^6 + \ldots \end{aligned}$$

and we can see a few patterns right away. The denominators are $j! = (2\ell)!$ and the numerators have factors of $(-2)^{j/2} = (-2)^{\ell}$. The factors of $n\,(n-2)\,(n-4)\ldots$ kind of look like factorials, but they change by two rather than one and they're also missing a few. If we pull out a 2 from each of the $j/2 = \ell$ factors, then

$$n\,(n-2)\,(n-4)\ldots = 2^{\ell}\left(\frac{n}{2}\right)\left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\ldots$$

However, we're shy $(n/2-\ell)$ factors of this being $(n/2)!$, so

$$n\,(n-2)\,(n-4)\ldots = 2^{\ell}\frac{\left(\frac{n}{2}\right)!}{\left(\frac{n}{2}-\ell\right)!}.$$

Combining all of this into a coefficient, we get

$$a_{2\ell} = \left[\frac{(-2)^{\ell}}{(2\ell)!}\right]\left[\frac{2^{\ell}\left(\frac{n}{2}\right)!}{\left(\frac{n}{2}-\ell\right)!}\right] a_0 = \frac{(-1)^{\ell}\,2^{2\ell}\left(\frac{n}{2}\right)!}{(2\ell)!\left(\frac{n}{2}-\ell\right)!}\,a_0$$

and, therefore,

$$u_{\text{even}} = \sum_{\ell=0}^{n/2}\left[\frac{(-1)^{\ell}\,2^{2\ell}\left(\frac{n}{2}\right)!}{(2\ell)!\left(\frac{n}{2}-\ell\right)!}\right] a_0\chi^{2\ell}$$

$$u_{\text{even}} = a_0 \left(\frac{n}{2}\right)! \sum_{\ell=0}^{n/2} \frac{(-1)^\ell}{(2\ell)!\left(\frac{n}{2}-\ell\right)!} (2\chi)^{2\ell}, \tag{9.4.74}$$

where a_0 is the normalization constant.

- The sum in Eq. 9.4.74 looks a lot like an even **Hermite polynomial**, which are given by

$$H_{\text{even}} = n! \sum_{\ell=0}^{n/2} \frac{(-1)^{\ell-\frac{n}{2}}}{(2\ell)!\left(\frac{n}{2}-\ell\right)!} (2\chi)^{2\ell}. \tag{9.4.75}$$

In fact, you could say

$$u_{\text{even}} = a_0 \frac{\left(\frac{n}{2}\right)!}{n!} (-1)^{n/2} H_n,$$

or, since we can just merge constants into the currently unknown a_0,

$$u_{\text{even}} = a_0 H_n. \tag{9.4.76}$$

That's what I call a simple solution! The odd solutions work out in a similar fashion, where odd Hermite polynomial's are given by

$$H_{\text{odd}} = n! \sum_{\ell=0}^{(n-1)/2} \frac{(-1)^{\ell-\frac{n-1}{2}}}{(2\ell+1)!\left(\frac{n-1}{2}-\ell\right)!} (2\chi)^{2\ell+1}. \tag{9.4.77}$$

and u is

$$u_{\text{odd}} = a_1 H_n. \tag{9.4.78}$$

The first ten Hermite polynomials are shown in Table 9.2.

- By Eq. 9.4.68, the eigenstates are

$$\Psi_n = u_n e^{-\chi/2} = C_n H_n e^{-\chi^2/2} \tag{9.4.79}$$

for odd and even values of n. We've combined them by replacing a_0 and a_1 with C_n since you never see both at the same time anyway. We

Table 9.2: This is the first ten Hermite polynomials, $H_n(\chi)$. They represent solutions to the harmonic oscillator in Example 9.4.6.

H_n		Hermite Polynomials
H_0	=	1
H_1	=	2χ
H_2	=	$4\chi^2 - 2$
H_3	=	$8\chi^3 - 12\chi$
H_4	=	$16\chi^4 - 48\chi^2 + 12$
H_5	=	$32\chi^5 - 160\chi^3 + 120\chi$
H_6	=	$64\chi^6 - 480\chi^4 + 720\chi^2 - 120$
H_7	=	$128\chi^7 - 1{,}344\chi^5 + 3{,}360\chi^3 - 1{,}680\chi$
H_8	=	$256\chi^8 - 3{,}584\chi^6 + 13{,}440\chi^4 - 13{,}440\chi^2 + 1{,}680$
H_9	=	$512\chi^9 - 9{,}216\chi^7 + 48{,}384\chi^5 - 80{,}640\chi^3 + 3{,}0240\chi$

still need to normalize to find C_n, which we can't do without a property of Hermite polynomials:

$$H_{n+1} - 2\chi H_n + 2nH_{n-1} = 0, \tag{9.4.80}$$

a recursive relation. We can justify this be substituting in Eqs. 9.4.75 and 9.4.77. Assuming H_n is even, then H_{n+1} and H_{n-1} will both be odd. Therefore,

$$\left\{ \begin{array}{rcl} H_{n+1} & = & (n+1)! \sum_{\ell=0}^{(n+1-1)/2} \frac{(-1)^{\ell-\frac{n+1-1}{2}}}{(2\ell+1)!\left(\frac{n+1-1}{2}-\ell\right)!} (2\chi)^{2\ell+1} \\ -2\chi H_n & = & -2\chi \left[n! \sum_{\ell=0}^{n/2} \frac{(-1)^{\ell-\frac{n}{2}}}{(2\ell)!\left(\frac{n}{2}-\ell\right)!} (2\chi)^{2\ell} \right] \\ 2nH_{n-1} & = & 2n \left[(n-1)! \sum_{\ell=0}^{(n-1-1)/2} \frac{(-1)^{\ell-\frac{n-1-1}{2}}}{(2\ell+1)!\left(\frac{n-1-1}{2}-\ell\right)!} (2\chi)^{2\ell+1} \right], \end{array} \right\}$$

$$\left\{ \begin{array}{rcl} H_{n+1} & = & (n+1)! \sum_{\ell=0}^{n/2} \frac{(-1)^{\ell-\frac{n}{2}}}{(2\ell+1)!\left(\frac{n}{2}-\ell\right)!} (2\chi)^{2\ell+1} \\ -2\chi H_n & = & -2\chi \left[n! \sum_{\ell=0}^{n/2} \frac{(-1)^{\ell-\frac{n}{2}}}{(2\ell)!\left(\frac{n}{2}-\ell\right)!} (2\chi)^{2\ell} \right] \\ 2nH_{n-1} & = & 2n \left[(n-1)! \sum_{\ell=0}^{\frac{n}{2}-1} \frac{(-1)^{\ell-\frac{n}{2}-1}}{(2\ell+1)!\left(\frac{n}{2}-1-\ell\right)!} (2\chi)^{2\ell+1} \right] \end{array} \right\},$$

where we've changed the n in each sum appropriately. If we play with the factorials a bit and move some things around, then

$$\left\{\begin{array}{rcl} H_{n+1} &=& n!\sum_{\ell=0}^{n/2}\frac{(n+1)(-1)^{\ell-\frac{n}{2}}}{(2\ell+1)!\left(\frac{n}{2}-\ell\right)!}(2\chi)^{2\ell+1} \\ -2\chi H_n &=& n!\sum_{\ell=0}^{n/2}\frac{-(2\ell+1)(-1)^{\ell-\frac{n}{2}}}{(2\ell+1)!\left(\frac{n}{2}-\ell\right)!}(2\chi)^{2\ell+1} \\ 2nH_{n-1} &=& n!\sum_{\ell=0}^{\frac{n}{2}-1}\frac{-2(-1)^{\ell-\frac{n}{2}}\left(\frac{n}{2}-\ell\right)}{(2\ell+1)!\left(\frac{n}{2}-\ell\right)!}(2\chi)^{2\ell+1} \end{array}\right\}.$$

We can add an $\ell = n/2$ term to the last sum because the factor $\left(\frac{n}{2}-\ell\right)$ would be zero anyway (i.e. *voodoo math*). Now that the limits on the sums are the same, we can add all three together and we get

$$n!\sum_{\ell=0}^{n/2}\left[(n+1)-(2\ell+1)-2\left(\frac{n}{2}-\ell\right)\right]\frac{(-1)^{\ell-\frac{n}{2}}}{(2\ell+1)!\left(\frac{n}{2}-\ell\right)!}(2\chi)^{2\ell+1}$$

and the quantity in square brackets is

$$(n+1)-(2\ell+1)-2\left(\tfrac{n}{2}-\ell\right) = n+1-2\ell-1-n+2\ell = 0,$$

which justifies Eq. 9.4.80.

- If we use two versions of Eq. 9.4.80,

$$\left\{\begin{array}{rcl} H_{n+1}-2\chi H_n+2nH_{n-1} &=& 0 \\ H_n-2\chi H_{n-1}+2nH_{n-2} &=& 0 \end{array}\right\},$$

and multiply by $-H_{n-1}$ and $+H_n$ respectively, then we get

$$\left\{\begin{array}{rcl} -H_{n+1}H_{n-1}+2\chi H_nH_{n-1}-2n\left(H_{n-1}\right)^2 &=& 0 \\ \left(H_n\right)^2-2\chi H_{n-1}H_n+2nH_{n-2}H_n &=& 0 \end{array}\right\}.$$

Now we can add these two together,

$$\left(H_n\right)^2-2n\left(H_{n-1}\right)^2+2nH_{n-2}H_n-H_{n+1}H_{n-1}=0,$$

and multiply everything by an arbitrary function, $f(\chi)$, to get

$$\left(H_n\right)^2 f(\chi)-2n\left(H_{n-1}\right)^2 f(\chi)+2nH_{n-2}H_nf(\chi)-H_{n+1}H_{n-1}f(\chi)=0.$$

If we integrate over all space, then the last two terms will go to zero because Hermite polynomials are orthogonal functions (they have to be if they're eigenfunctions). This gives us

$$\int_{-\infty}^{+\infty} (H_n)^2 f(\chi)\ d\chi - 2n \int_{-\infty}^{+\infty} (H_{n-1})^2 f(\chi)\ d\chi = 0$$

$$\int_{-\infty}^{+\infty} (H_n)^2 f(\chi)\ d\chi = 2n \int_{-\infty}^{+\infty} (H_{n-1})^2 f(\chi)\ d\chi,$$

which is still recursive. However, if we perform this operation enough times,

$$\begin{aligned} \int_{-\infty}^{+\infty} (H_n)^2 f(\chi)\ d\chi &= 2^2 n\,(n-1) \int_{-\infty}^{+\infty} (H_{n-2})^2 f(\chi)\ d\chi \\ &= 2^3 n\,(n-1)\,(n-2) \int_{-\infty}^{+\infty} (H_{n-3})^2 f(\chi)\ d\chi \\ &= 2^4 n\,(n-1)\,(n-2)\,(n-3)\ldots, \end{aligned}$$

then *eventually* we'll get to $H_0 = 1$,

$$\int_{-\infty}^{+\infty} (H_n)^2 f(\chi)\ d\chi = 2^n n! \int_{-\infty}^{+\infty} f(\chi)\ d\chi\,, \tag{9.4.81}$$

which is exactly what we need to normalize Ψ (Eq. 9.4.79).

- Using the normalization condition (Eq. 9.2.17), the fact that

$$\psi^*\psi = \left(\Psi^* e^{iEt/\hbar}\right)\left(\Psi e^{-iEt/\hbar}\right) = \Psi^*\Psi,$$

and that Ψ is entirely real, we get

$$\int_{-\infty}^{+\infty} \Psi^*\Psi\ dx = \int_{-\infty}^{+\infty} (\Psi)^2\, dx = 1.$$

Notice the use of x rather than χ? That's important. Using the chain rule for derivatives (Eq. 3.1.2) and definition of χ (Eq. 9.4.64), this is actually

$$\int_{-\infty}^{+\infty} (\Psi)^2 \frac{dx}{d\chi}\, d\chi = \int_{-\infty}^{+\infty} (\Psi)^2 \sqrt{\tfrac{\hbar}{m\omega}}\, d\chi = \sqrt{\tfrac{\hbar}{m\omega}} \int_{-\infty}^{+\infty} (\Psi)^2\, d\chi = 1$$

Now we're in a position to be using functions of χ. Substituting from Eq. 9.4.79, we get

$$\sqrt{\tfrac{\hbar}{m\omega}}\int_{-\infty}^{+\infty}\left(C_n H_n e^{-\chi^2/2}\right)^2 d\chi = 1$$

$$C_n^2\sqrt{\tfrac{\hbar}{m\omega}}\int_{-\infty}^{+\infty}\left(H_n\right)^2 e^{-\chi^2} d\chi = 1.$$

This integral matches Eq. 9.4.81 if we set $f = e^{-\chi^2}$, so

$$C_n^2\sqrt{\tfrac{\hbar}{m\omega}}\,(2^n n!)\int_{-\infty}^{+\infty} e^{-\chi^2} d\chi = 1$$

and we just have to solve the remaining integral.

- In its current state, it's unsolveable. However, if we perform a little *voodoo math* (with a little foresight; we can add zeros, multiply by ones, add and subtract constants, etc. to simplify a mathematical expression). This time we're going to multiple the integral by itself and then square root. Using more familiar symbols for now,

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\int_{-\infty}^{+\infty} e^{-x^2} dx \int_{-\infty}^{+\infty} e^{-y^2} dy},$$

where we've changed the variable in the second integral. Combining the integrals, we get

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(x^2+y^2)} dx\, dy}.$$

This is a double integral that we can easily transform to polar coordinates (see Section 1.2). The result is

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\int_0^{2\pi}\int_0^{+\infty} e^{-s^2} s\, ds\, d\phi},$$

where the extra factor of s appears due to the Jacobian (Eq. 6.6.1). A quick u-substitution of $u = s^2$ (and $du = 2s\, ds$) gives us

$$\sqrt{\int_0^{2\pi}\int_0^{+\infty} e^{-u}\tfrac{1}{2} du\, d\phi} = \sqrt{\int_0^{2\pi}\tfrac{1}{2} d\phi} = \sqrt{\tfrac{1}{2}\int_0^{2\pi} d\phi} = \sqrt{\tfrac{1}{2}\,(2\pi)}$$

or just

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}. \tag{9.4.82}$$

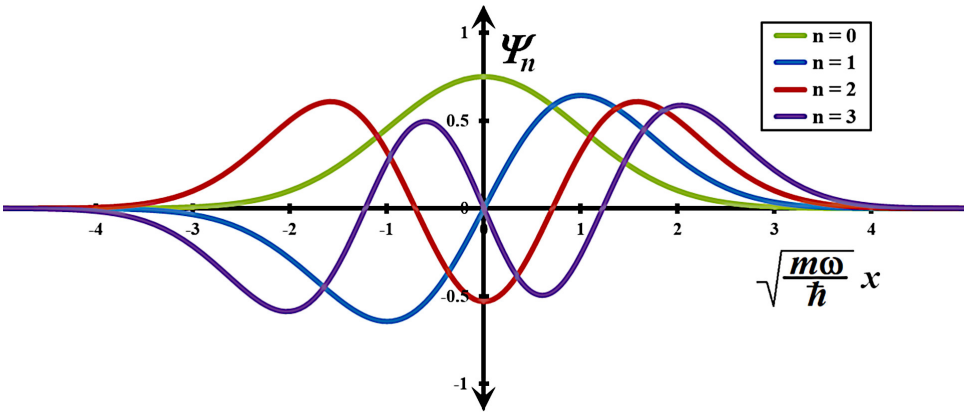


Figure 9.19: These are the first four eigenstates in the finite square well given in Example 9.4.6.

- Therefore, the normalization constant is

$$C_n^2\sqrt{\frac{\hbar}{m\omega}}\,(2^n n!)\,\sqrt{\pi} = 1 \quad \Rightarrow C_n^2 = \sqrt{\frac{m\omega}{\hbar\pi}}\frac{1}{2^n n!}$$

$$\Rightarrow C_n = \left(\frac{m\omega}{\hbar\pi}\right)^{1/4}\frac{1}{\sqrt{2^n n!}} \tag{9.4.83}$$

and the eigenstates are given by

$$\Psi_n(\chi) = \left(\frac{m\omega}{\hbar\pi}\right)^{1/4}\frac{1}{\sqrt{2^n n!}}\,H_n(\chi)\;e^{-\chi^2/2}. \tag{9.4.84}$$

Transforming this back to x using Eq. 9.4.64, we get

$$\boxed{\Psi_n(x) = \left(\frac{m\omega}{\hbar\pi}\right)^{1/4}\frac{1}{\sqrt{2^n n!}}\left[H_n\left(\sqrt{\frac{m\omega}{\hbar}}\,x\right)\right]e^{-m\omega x^2/(2\hbar)}}, \tag{9.4.85}$$

where n is a whole number (i.e. $n = 1, 2, 3, \ldots$) representing the energy level (Eq. 9.4.72) and H_n is the appropriate Hermite polynomial (Table 9.2).

- Unfortunately, the eigenstates are only the stationary states at $t = 0$. In general, stationary states are given by Eq. 9.3.13, so

$$\psi_n(x,t) = \Psi_n(x)\,e^{-iEt/\hbar} = \Psi_n(x)\,e^{-i(n+\frac{1}{2})\omega t}$$

$$\boxed{\psi_n(x,t) = \left(\tfrac{m\omega}{\hbar\pi}\right)^{1/4} \tfrac{1}{\sqrt{2^n n!}} \left[H_n\left(\sqrt{\tfrac{m\omega}{\hbar}}\, x\right)\right] e^{-m\omega x^2/(2\hbar) - i(n+\frac{1}{2})\omega t}}, \quad (9.4.86)$$

where n is a whole number (i.e. $n = 1, 2, 3, \ldots$) representing the energy level (Eq. 9.4.72) and H_n is the appropriate Hermite polynomial (Table 9.2).

Example 9.4.6 only works this model out in one dimension, so you may be skeptical about its relevance to Planck's result (Eq. 9.4.61). However, the three-dimensional case (Eq. 9.4.59) works out just like it did for the infinite well (see Example 9.4.2). The differential equation separates into three independent equations (one of x, y, and z). Therefore, the energy levels are given by

$$E_{n_x n_y n_z} = E_{n_x} + E_{n_y} + E_{n_z}$$

$$E_{n_x n_y n_z} = \left(n_x + \frac{1}{2}\right)\hbar\omega + \left(n_y + \frac{1}{2}\right)\hbar\omega + \left(n_z + \frac{1}{2}\right)\hbar\omega$$

$$E_{n_x n_y n_z} = \left(n_x + n_y + n_z + \frac{3}{2}\right)\hbar\omega, \quad (9.4.87)$$

where n_x, n_y, and n_z are whole numbers (i.e. $n_i = 1, 2, 3, \ldots$). If we define $n = n_x + n_y + n_z$, then

$$E_n = \left(n + \frac{3}{2}\right)\hbar\omega, \quad (9.4.88)$$

where n is a whole number (i.e. $n = 1, 2, 3, \ldots$).

There is some degeneracy (i.e. multiple states having the same energy) just like with the three-dimensional infinite well (see Example 9.4.2), but that has no effect on our ultimate point. Let's say an electron transitions from a higher stationary state (n_i) to a lower one (n_f). The loss of energy is

$$-\Delta E = E_i - E_f = (n_i - n_f)\,\hbar\omega,$$

where $(n_i - n_f)$ is a whole number (i.e. $1, 2, 3, \ldots$). This means the loss of energy is a whole number multiple of $\hbar\omega$, which we know is emitted as a photon. Planck's result (Eq. 9.4.61) is supported even in the three-dimensional case.

Chapter 10

Modern Quantum Mechanics

10.1 Finding Wave Functions

In Section 9.4, we saw some simple models with very narrow applications. Even so, the methods used in developing those models are the same as those used for more realistic ones. We've seen a lot of methodical processes in this book and it should be no surprise there is one for finding stationary states:

1. *Determine the potential energy function for the system.* This is what makes models different from one another.

2. *Plug this into the time-independent Schrödinger equation (Eq. 9.3.14).* It's easier to find the eigenstates before you find stationary states.

3. *Use methods from differential equations to solve for the eigenstates.* This may differ depending on the appearance of the differential equation, but common solutions usually involve sine, cosine, the exponential, and/or famous sets of polynomials.

4. *Apply boundary conditions to find any unknowns that may have appeared in the last step.* **The eigenstates and their derivatives must be continuous and finite.**

5. *Normalize using Eq. 9.2.17 to find the one remaining unknown: the normalization constant.* If you have more than one unknown left at this step, then you didn't finish the last step.

6. *Multiply by the time-evolution factor to find the stationary states.* Remember, Eq. 9.3.13?

A lot of the details inside these steps are the same as well, so we're going to do a lot of referencing (and step skipping) in this section to save time. I don't want this book to look like it's just on quantum mechanics.

10.2 Single-Electron Atoms

Most atoms have many electrons, which poses all sorts of mathematical difficulties (the so called "three-body problem"). We'll discuss these difficulties in Section 10.3. For now, I think it's best to stick with single-electron atoms. Hydrogen is an obvious example, but the results we'll be getting will also apply to entities like singly-ionized helium and doubly-ionized lithium since this says nothing of the nucleus. As far as the electron is concerned, the nucleus is just one particle.

If we hope to model *real* atoms, then we need to be applying electrodynamics (Chapter 5) since that's what holds the atom together. We're going to make two assumptions to keep things simple:

1. The nucleus is very small in size compared the atom, which is *always* accurate since $r_{\text{atom}} \approx (10^4 \text{ to } 10^5)\, r_{\text{nuc}}$.

2. The nucleus is stationary, which is accurate if $m_{\text{nuc}} \gg m_e$.

Given these, the potential energy of the electron should be related to Coulomb's law (Eq. 5.2.1) for point charges. Force is related to potential energy by $\vec{F} = -\vec{\nabla} V$ (Eq. 4.2.3), so we can say

$$\vec{F}_E = -\left(\frac{\partial V}{\partial r}\hat{r} + \frac{1}{r}\frac{\partial V}{\partial \theta}\hat{\theta} + \frac{1}{r\sin\theta}\frac{\partial V}{\partial \phi}\hat{\phi} \right)$$

in spherical coordinates (Eq. 3.3.6) for the atom. Substituting in Coulomb's law (Eq. 5.2.1), we get

$$k_E \frac{q_1 q_2}{r^2}\hat{r} = -\frac{\partial V}{\partial r}\hat{r} - \frac{1}{r}\frac{\partial V}{\partial \theta}\hat{\theta} - \frac{1}{r\sin\theta}\frac{\partial V}{\partial \phi}\hat{\phi}$$

$$\Rightarrow \left\{ \; k_E \frac{q_1 q_2}{r^2} = -\frac{\partial V}{\partial r} \quad , \quad 0 = \frac{\partial V}{\partial \theta} \quad , \quad 0 = \frac{\partial V}{\partial \phi} \; \right\}$$

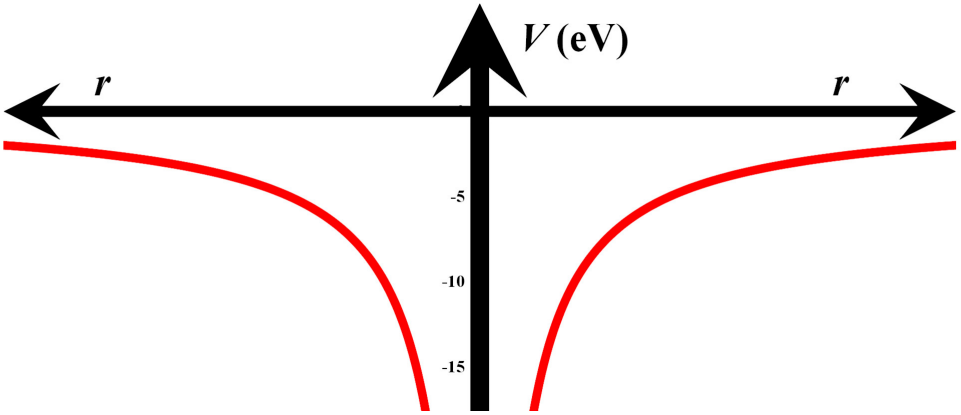


Figure 10.1: This is the potential energy function (Eq. 10.2.1) for the hydrogen atom where r represents the distance from the proton in the nucleus. Values on the vertical axis are in units of electron volts (eV).

by equating vector components. That means V only has radial dependence and integrating gives us

$$V(r) = \int_r^\infty k_E \frac{q_1 q_2}{r^2}\, dr = k_E \left[\frac{-q_1 q_2}{r}\right]_r^\infty = k_E \left[0 - \frac{-q_1 q_2}{r}\right] = k_E \frac{q_1 q_2}{r}$$

where we've defined V traditionally (i.e. $V \to 0$ as $r \to \infty$).

In the atom, one of the charges is an electron (i.e. $q_1 = q_e = -q$) and the other is the arbitrary nucleus containing Z protons (i.e. $q_2 = Zq_p = Zq$), where $q = q_p = 1.602 \times 10^{-19}$ C is the **elementary charge**. This simplifies the potential energy function to

$$\boxed{V(r) = -k_E \frac{Zq^2}{r} = -\frac{Zq^2}{4\pi\epsilon_0 r}}, \tag{10.2.1}$$

where $\epsilon_0 = (4\pi k_E)^{-1} = 8.854 \times 10^{-12}$ C^2/(Nm^2) is just the permittivity of free space from Chapter 5.

Example 10.2.1

What are the stationary states (and corresponding energies) for a *lone* non-relativistic electron bound by a positive nucleus?

- The time-independent Schrödinger equation (Eq. 9.3.14) will take the form

$$\mathcal{H}\Psi = -\frac{\hbar^2}{2m}\vec{\nabla}^2\Psi + V\Psi = E\Psi$$

$$-\frac{\hbar^2}{2m}\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Psi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial\Psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\Psi}{\partial\phi^2}\right] - \frac{Zq^2}{4\pi\epsilon_0 r}\Psi = E\Psi \tag{10.2.2}$$

  where $\vec{\nabla}^2$ has been expanded for spherical coordinates (Eq. 3.3.9) and $m = m_e = 9.109\times10^{-31}$ kg. I know this looks pretty nasty, but we can handle it.

- Using separation of variables like we did in Example 9.4.2, we can say

$$\Psi(r,\theta,\phi) = R(r)\,\Theta(\theta)\,\Phi(\phi)\,. \tag{10.2.3}$$

  Eventually, we'll have to normalize Ψ, but we can save time by normalizing R, Θ, and Φ individually. We know

$$\psi^*\psi = \left(\Psi^* e^{iEt/\hbar}\right)\left(\Psi e^{-iEt/\hbar}\right) = \Psi^*\Psi,$$

  so normalization condition (Eq. 9.2.17) would be

$$\begin{aligned}
\int_0^{2\pi}\int_0^{\pi}\int_0^{\infty}\Psi^*\Psi\,r^2\sin\theta\,dr\,d\theta\,d\phi &= 1\\
\int_0^{2\pi}\int_0^{\pi}\int_0^{\infty}\left(R^*\Theta^*\Phi^*\right)\left(R\Theta\Phi\right)r^2\sin\theta\,dr\,d\theta\,d\phi &= 1\\
\left(\int_0^{\infty}R^*R\,r^2dr\right)\left(\int_0^{\pi}\Theta^*\Theta\,\sin\theta\,d\theta\right)\left(\int_0^{2\pi}\Phi^*\Phi\,d\phi\right) &= 1
\end{aligned}$$

  where the integrals are also separable. Since $(1)\,(1)\,(1) = 1$, we could just as easily say

$$\int_0^{\infty}R^*R\,r^2dr = 1 \tag{10.2.4a}$$

$$\int_0^{\pi}\Theta^*\Theta\,\sin\theta\,d\theta = 1 \tag{10.2.4b}$$

$$\int_0^{2\pi}\Phi^*\Phi\,d\phi = 1. \tag{10.2.4c}$$

- If we substitute Eq. 10.2.3 into 10.2.2, multiply through by $-\frac{2mr^2}{\hbar^2 R\Theta\Phi}$, and set it equal to zero, then

$$\frac{1}{R}\frac{\partial}{\partial r}\left(r^2\frac{\partial R}{\partial r}\right)+\frac{1}{\Theta\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial\Theta}{\partial\theta}\right)+\frac{1}{\Phi\sin^2\theta}\frac{\partial^2\Phi}{\partial\phi^2}+\frac{Zmq^2r}{2\pi\epsilon_0\hbar^2}+\frac{2mr^2}{\hbar^2}E=0.$$

We can see that all terms dependent on r are *not* dependent on θ or ϕ (and vice versa), so the only way their sum can *always* be zero is if each term is individually constant and they cancel each other. Therefore, this is just two independent differential equations:

$$\frac{1}{R}\frac{\partial}{\partial r}\left(r^2\frac{\partial R}{\partial r}\right)+\frac{Zmq^2r}{2\pi\epsilon_0\hbar^2}+\frac{2mr^2}{\hbar^2}E=\ell\,(\ell+1) \tag{10.2.5a}$$

$$\frac{1}{\Theta\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial\Theta}{\partial\theta}\right)+\frac{1}{\Phi\sin^2\theta}\frac{\partial^2\Phi}{\partial\phi^2}=-\ell\,(\ell+1) \tag{10.2.5b}$$

where ℓ is a constant. We could have chosen something simple like k, but $\ell\,(\ell+1)$ is going to make our lives easier in Eq. 10.2.5b.

- If we take Eq. 10.2.5b, multiply through by $\sin^2\theta$, and set it equal to zero, then

$$\frac{\sin\theta}{\Theta}\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial\Theta}{\partial\theta}\right)+\frac{1}{\Phi}\frac{\partial^2\Phi}{\partial\phi^2}+\ell\,(\ell+1)\sin^2\theta=0.$$

We can see that all terms dependent on θ are *not* dependent on ϕ (and vice versa), so the only way their sum can *always* be zero is if each term is individually constant and they cancel each other. Therefore, this is just two independent differential equations:

$$\frac{\sin\theta}{\Theta}\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial\Theta}{\partial\theta}\right)+\ell\,(\ell+1)\sin^2\theta = m_\ell^2 \tag{10.2.6a}$$

$$\frac{1}{\Phi}\frac{\partial^2\Phi}{\partial\phi^2} = -m_\ell^2 \tag{10.2.6b}$$

where m_ℓ is a constant. Again, we could have chosen something simple like k, but m_ℓ is going to make our lives easier in Eq. 10.2.6b. We've also given it an ℓ subscript to distinguish it from the mass m.

- Now we have three completely independent differential equations: Eq. 10.2.5a, 10.2.6a, and 10.2.6b (one for each variable). To summarize, lets multiply each by their portion of the eigenstate (R, Θ, and Φ, respectively) and move a couple terms around. They are now

$$\boxed{\frac{\partial}{\partial r}\left(r^2\frac{\partial R}{\partial r}\right)+\left[\frac{Zmq^2r}{2\pi\epsilon_0\hbar^2}+\frac{2mr^2}{\hbar^2}E-\ell\left(\ell+1\right)\right]R=0} \tag{10.2.7a}$$

$$\boxed{\left(\sin\theta\right)\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial\Theta}{\partial\theta}\right)+\left[\ell\left(\ell+1\right)\sin^2\theta-m_\ell^2\right]\Theta=0} \tag{10.2.7b}$$

$$\boxed{\frac{\partial^2\Phi}{\partial\phi^2}=-m_\ell^2\Phi} \tag{10.2.7c}$$

  and we need different methods to solve each one.

- Eq. 10.2.7c should be very familiar to you at this point. It appeared in both the infinite well (Example 9.4.1) and the finite well (Example 9.4.3). In those simple cases, we said the only two functions with second derivatives proportional to the negative of themselves were $\sin(m_\ell x)$ and $\cos(m_\ell x)$. This is true only if you're interested in *real* solutions. In general, *complex* solutions are permitted, so we should really apply Euler's formula,

$$e^{i\phi}=\cos\phi+i\sin\phi, \tag{10.2.8}$$

  and say $e^{im_\ell\phi}$ and $e^{-im_\ell\phi}$ are the functions instead. We did this before in deriving Eq. 9.2.1 without explicitly stating it. The general solution will be a linear combination of the two, so

$$\Phi(\phi)=C_1e^{im_\ell\phi}+C_2e^{-im_\ell\phi},$$

  where C_1 and C_2 are just constants. We can simplify further by merging the sign in the exponent with m_ℓ, which results in

$$\boxed{\Phi^{m_\ell}(\phi)=C_\phi\,e^{im_\ell\phi}=\frac{1}{\sqrt{2\pi}}\,e^{im_\ell\phi}}, \tag{10.2.9}$$

  where m_ℓ acts as a label (*not* an exponent) on the variable Φ^{m_ℓ}. This is similar to contravariant indices from Chapter 6, but raising and lowering is irrelevant because Φ^{m_ℓ} is *not* a rank-1 tensor. The constant $C_\phi=1/\sqrt{2\pi}$ was determined by normalizing using Eq. 10.2.4c.

- However, because ϕ is the azimuthal angle, we know

$$\Phi(0) = \Phi(2\pi) \quad \Rightarrow 1 = e^{i\,2\pi m_\ell}.$$

By Euler's formula (Eq. 10.2.8), this means

$$1 + i\,0 = \cos(2\pi m_\ell) + i\sin(2\pi m_\ell) \quad \Rightarrow 0 = \sin(2\pi m_\ell)$$

implying that m_ℓ must be an integer (i.e. $m_\ell = 0, \pm 1, \pm 2, \ldots$).

- Eq. 10.2.7b probably doesn't look familiar, but it's called a **generalized Legendre equation** which are traditionally defined as

$$\frac{d}{dx}\left[\left(1-x^2\right)\frac{dy}{dx}\right] + \left[\ell\,(\ell+1) - \frac{m_\ell^2}{1-x^2}\right] y = 0. \tag{10.2.10}$$

Still skeptical? In our case, $x = \cos\theta$, so $dx = -\sin\theta\, d\theta$ and $(1-x^2) = \sin^2\theta$. Substituting these into Eq. 10.2.10, we get

$$\begin{aligned}
\frac{1}{-\sin\theta}\frac{\partial}{\partial\theta}\left[\sin^2\theta\frac{1}{-\sin\theta}\frac{\partial\Theta}{\partial\theta}\right] + \left[\ell\,(\ell+1) - \frac{m_\ell^2}{\sin^2\theta}\right]\Theta &= 0 \\
\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left[\sin\theta\frac{\partial\Theta}{\partial\theta}\right] + \left[\ell\,(\ell+1) - \frac{m_\ell^2}{\sin^2\theta}\right]\Theta &= 0
\end{aligned}$$

and multiplying through by $\sin^2\theta$ results in Eq. 10.2.7b. Solutions to Eq. 10.2.10 are called **associated Legendre functions**,

$$P_\ell^{m_\ell}(x) = \frac{(-1)^{m_\ell}}{2^\ell \ell!}\left(1-x^2\right)^{m_\ell/2}\frac{d^{\ell+m_\ell}}{dx^{\ell+m_\ell}}\left(x^2-1\right)^\ell, \tag{10.2.11}$$

where ℓ and m_ℓ act as labels (*not* as exponents) on the variable $P_\ell^{m_\ell}$. This is similar to contravariant indices from Chapter 6, but raising and lowering is irrelevant because $P_\ell^{m_\ell}$ is *not* a rank-2 tensor. Functions for negative m_ℓ values are related to their positive counterpart by

$$P_\ell^{-m_\ell} = (-1)^{m_\ell}\frac{(\ell-m_\ell)!}{(\ell+m_\ell)!}, P_\ell^{m_\ell} \tag{10.2.12}$$

which might save you some time. If these functions are solutions to Eq. 10.2.7b, then the general solutions take the form

$$\Theta_\ell^{m_\ell}(\theta) = C_\theta\, P_\ell^{m_\ell}(\cos\theta)$$

Table 10.1: This is the first ten associated Legendre functions, $P_\ell^{m_\ell}(x) = P_\ell^{m_\ell}(\cos\theta)$. They represent solutions for $\Theta(\theta)$ in Example 10.2.1. Note: Functions for negative m_ℓ are given by Eq. 10.2.12.

$P_\ell^{m_\ell}$		Legendre Functions		
P_0^0	$=$	1	$=$	1
P_1^0	$=$	x	$=$	$\cos\theta$
P_1^1	$=$	$-\sqrt{1-x^2}$	$=$	$-\sin\theta$
P_2^0	$=$	$\frac{1}{2}(3x^2-1)$	$=$	$\frac{1}{2}(3\cos^2\theta-1)$
P_2^1	$=$	$-3x\sqrt{1-x^2}$	$=$	$-3\cos\theta\,\sin\theta$
P_2^2	$=$	$3(1-x^2)$	$=$	$3\sin^2\theta$
P_3^0	$=$	$\frac{1}{2}(5x^3-3x)$	$=$	$\frac{1}{2}(5\cos^3\theta-3\cos\theta)$
P_3^1	$=$	$-\frac{3}{2}(5x^2-1)\sqrt{1-x^2}$	$=$	$-\frac{3}{2}(5\cos^2\theta-1)\sin\theta$
P_3^2	$=$	$15x(1-x^2)$	$=$	$15\cos\theta\,\sin^2\theta$
P_3^3	$=$	$-15(1-x^2)^{3/2}$	$=$	$-15\sin^3\theta$

$$\boxed{\Theta_\ell^{m_\ell}(\theta) = \sqrt{\frac{(2\ell+1)}{2}\frac{(\ell-m_\ell)!}{(\ell+m_\ell)!}}\,P_\ell^{m_\ell}(\cos\theta)}, \tag{10.2.13}$$

where ℓ and m_ℓ act as labels (*not* as exponents) on the variables $\Theta_\ell^{m_\ell}$ just like they do for $P_\ell^{m_\ell}$. The constant C_θ was determined by normalizing using Eq. 10.2.4b (and some properties of Legendre functions).

- These are called "functions" rather than "polynomials" because, if m_ℓ is odd, then $P_\ell^{m_\ell}(x)$ contains a factor of $\sqrt{1-x^2}$. In this example, $(1-x^2) = \sin^2\theta$ eliminating any square roots, but now you have trigonometric functions and it's still *technically* not a polynomial. I could write Eq. 10.2.11 in terms of θ, but that's a bit cumbersome in the derivatives. It's easier to find the function in terms of x, then use $x = \cos\theta$ and $(1-x^2) = \sin^2\theta$ to transform. I've provided a list in Table 10.1.

- Eq. 10.2.11 implies ℓ must be a whole number (i.e. $\ell = 0, 1, 2, 3, \ldots$). We've already seen that m_ℓ must be an integer, but now it has limits dictated by ℓ: $|m_\ell| \le \ell$ or

$$m_\ell = -\ell,\ -\ell+1,\ \ldots,\ \ell-1,\ \ell. \tag{10.2.14}$$

These limits occur because both ℓ and m_ℓ count the number of derivatives in Eq. 10.2.11. If $m_\ell > \ell$, then $P_\ell^{m_\ell} = 0$. If $m_\ell < -\ell$, then you have a negative number of derivatives and that doesn't make sense.

- The radial equation (Eq. 10.2.7a) is much more tedious than the other two, so we're going to work through this as succinctly as possible. Just like for the harmonic oscillator (see Example 9.4.6), we'll simplify the process by changing to a unitless variable:

$$\rho \equiv \frac{mq^2 r}{4\pi\hbar^2\epsilon_0} = \frac{r}{a_0} \tag{10.2.15}$$

where

$$a_0 \equiv \frac{4\pi\hbar^2\epsilon_0}{mq^2} = 0.0529 \text{ nm} \tag{10.2.16}$$

is the **Bohr radius** (see Eq. 9.1.2). With a little foresight from Eq. 9.1.3, we can even define a constant n such that

$$\frac{2m}{\hbar^2}E = -\frac{Z^2}{n^2 a_0^2}. \tag{10.2.17}$$

We already know E should be negative because V is negative *everywhere*. However, it's important to recognize all we know about n is that it's a *real* number. Any further restrictions (like saying it's a *natural* number) must be proven. Using Eq. 10.2.15, Eq. 10.2.17, and the product rule for derivatives (Eq. 3.1.5) on the radial equation (Eq. 10.2.7a), we get

$$\rho^2\frac{\partial^2 R}{\partial\rho^2} + 2\rho\frac{\partial R}{\partial\rho} + \left[2Z\rho - \frac{Z^2}{n^2}\rho^2 - \ell\,(\ell+1)\right] R = 0 \tag{10.2.18}$$

- Also, just like in Example 9.4.6, we can pull out factors to account for end-behavior because we know what the function should look like there. As $\rho \to \infty$, the radial equation approaches

$$\frac{\partial^2 R}{\partial\rho^2} - \frac{Z^2}{n^2}\,R \approx 0$$

because the ρ^2 terms dominate (i.e. they get bigger faster). That means $R(\rho)$ should include a factor of $e^{-Z\rho/n}$. Technically, $e^{+Z\rho/n}$ is also a

solution, but we ignored it since R must be finite. As $\rho \to 0$, the radial equation approaches

$$\rho^2 \frac{\partial^2 R}{\partial \rho^2} - \ell\,(\ell+1)\ R \approx 0,$$

where we've kept the second derivative term because it we're trying to avoid trivial solutions. This equation may be less familiar to you, but it's called a **Bernoulli differential equation** and its solutions are in the form ρ^α. In this case, they're ρ^ℓ and $\rho^{-(\ell+1)}$. However, $\rho^{-(\ell+1)} \to \infty$ as $\rho \to 0$, so only ρ^ℓ is a viable solution and $R(\rho)$ should include it as a factor. If we call everything else u, then

$$R(\rho) = \rho^\ell\ u(\rho)\ e^{-Z\rho/n} \tag{10.2.19}$$

and now we only need to determine the form of u.

- We could use a power series solution like we did for the harmonic oscillator (see Example 9.4.6), but that's extremely long and we can do better. If we substitute Eq. 10.2.19 into our new radial equation (Eq. 10.2.18), then we end up with a 15-term differential equation. Some of those terms either group or cancel leaving us with a 7-term differential equation. A couple pages of math later, we get

$$\rho \frac{\partial^2 u}{\partial \rho^2} + \left[(2\ell+1) + 1 - \left(\frac{2Z}{n}\rho\right)\right] \frac{\partial u}{\partial \rho} + \frac{2Z}{n}\,[n-\ell-1]\,u = 0$$

or

$$\left(\frac{2Z}{n}\rho\right) \frac{\partial^2 u}{\partial\,(2Z\rho/n)^2} + \left[(2\ell+1) + 1 - \left(\frac{2Z}{n}\rho\right)\right] \frac{\partial u}{\partial\,(2Z\rho/n)} + [n-\ell-1]\,u = 0.$$

This is called a **generalized Laguerre equation** defined as

$$x\frac{\partial^2 y}{\partial x^2} + (\alpha + 1 - x)\frac{\partial y}{\partial x} + \beta y = 0 \tag{10.2.20}$$

and its solutions are called **associated Laguerre polynomials** defined by

$$L_\beta^\alpha(x) = \frac{x^{-\alpha}e^x}{\beta!}\frac{d^\beta}{dx^\beta}\left(e^{-x}x^{\beta+\alpha}\right), \tag{10.2.21}$$

Table 10.2: This is the first several associated Laguerre polynomials, $L_\beta^\alpha(x)$. The value of α has been left open for the sake of generality. They represent solutions for $R(r)$ in Example 10.2.1.

L_β^α		Laguerre Polynomials
L_0^α	$=$	1
L_1^α	$=$	$1+\alpha-x$
L_2^α	$=$	$\frac{1}{2}\left[(\alpha+1)(\alpha+2)-2(\alpha+2)x+x^2\right]$
L_3^α	$=$	$\frac{1}{6}\left[(\alpha+1)(\alpha+2)(\alpha+3)-3(\alpha+2)(\alpha+3)x+3(\alpha+3)x^2-x^3\right]$

where α and β act as labels (*not* as exponents) on the variable L_β^α. This is similar to contravariant indices from Chapter 6, but raising and lowering is irrelevant because L_β^α is *not* a rank-2 tensor. However, both α and β must be whole numbers (i.e. $0, 1, 2, 3, \ldots$). I've provided a list in Table 10.1.

- In this example, $x = 2Z\rho/n$, $\alpha = 2\ell+1$, and $\beta = n-\ell-1$; so

$$u(\rho) = C_r\, L_{n-\ell-1}^{2\ell+1}\left(\frac{2Z}{n}\rho\right)$$

and, by Eq. 10.2.19,

$$R(\rho) = C_r\rho^\ell\left[L_{n-\ell-1}^{2\ell+1}\left(\frac{2Z}{n}\rho\right)\right]e^{-Z\rho/n}.$$

Using Eq. 10.2.15 to transform back to r, we get

$$\boxed{R_{n\ell}(r) = C_r\left(\frac{r}{a_0}\right)^\ell\left[L_{n-\ell-1}^{2\ell+1}\left(\frac{2Z}{n}\frac{r}{a_0}\right)\right]e^{-Zr/(na_0)}} \tag{10.2.22}$$

and, using Eq. 10.2.4a (and some properties of Laguerre polynomials) to normalize,

$$\boxed{C_r = a_0^{-3/2}\sqrt{\left(\frac{2Z}{n}\right)^{2\ell+3}\frac{(n-\ell-1)!}{2n\,(n+\ell)!}}}. \tag{10.2.23}$$

Some examples for the hydrogen atom ($Z = 1$) are shown in Table 10.4.

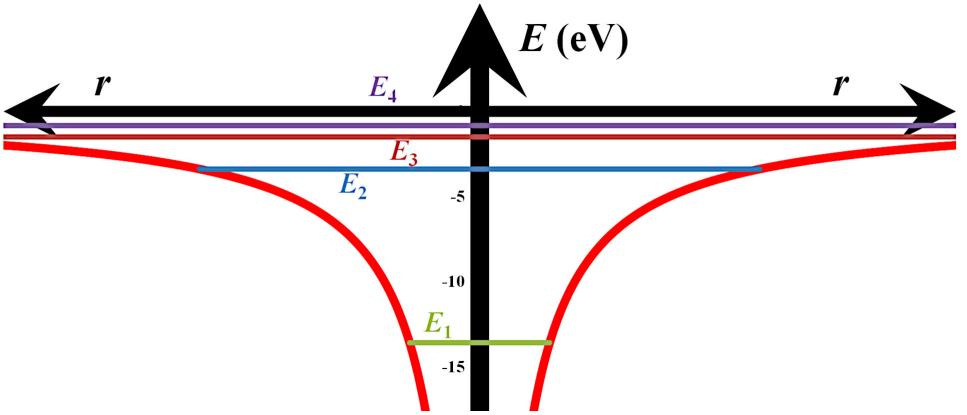


Figure 10.2: This is an **energy level diagram** showing the first four energy states for the hydrogen atom (i.e. for Example 10.2.1 with $Z = 1$) where r represents the distance from the proton in the nucleus.. The colors match those used in Figures 10.3 and 10.4.

- A consequence of Eq. 10.2.22 is that, since $(n - \ell - 1)$ must be a whole number, we have new restrictions on n and ℓ. First, n must be a natural number:

$$n = 1,\ 2,\ 3,\ 4,\ \ldots, \tag{10.2.24}$$

which is exactly what we expected. By Eq. 10.2.17, the energy levels are given by

$$\boxed{E_n = -\frac{Z^2}{n^2}\frac{\hbar^2}{2ma_0^2} = \frac{Z^2}{n^2}\,(-13.6\text{ eV})} \tag{10.2.25}$$

which matches Eq. 9.1.3. Also, we know $\ell < n$ or

$$\ell = 0,\ 1,\ \ldots,\ n - 1. \tag{10.2.26}$$

- To find the full eigenstate, we need to combine the parts using Eq. 10.2.3. Using a consistent labeling scheme to keep you from seeing any implied summations, that is

$$\Psi_{n\ell}^{m_\ell} = R_{n\ell}\,\Theta_\ell^{m_\ell}\,\Phi^{m_\ell}.$$

Substituting from Eqs. 10.2.22, 10.2.13, and 10.2.9 and we get

$$\boxed{\Psi_{n\ell}^{m_\ell} = C\left(\frac{r}{a_0}\right)^{\ell}\left[L_{n-\ell-1}^{2\ell+1}\left(\frac{2Z}{n}\frac{r}{a_0}\right)\right][P_\ell^{m_\ell}(\cos\theta)]\,e^{-Zr/(na_0)+im_\ell\phi}} \quad (10.2.27)$$

where

$$\boxed{C = a_0^{-3/2}\sqrt{\left(\frac{2Z}{n}\right)^{2\ell+3}\frac{(n-\ell-1)!}{2n\,(n+\ell)!}\frac{(2\ell+1)}{4\pi}\frac{(\ell-m_\ell)!}{(\ell+m_\ell)!}}}. \quad (10.2.28)$$

The quantity m_ℓ is acting act as a label (*not* an exponent) on the variable $\Psi_{n\ell}^{m_\ell}$. This is similar to contravariant indices from Chapter 6, but raising and lowering is irrelevant because $\Psi_{n\ell}^{m_\ell}$ is *not* a rank-3 tensor. The possible values for n, ℓ, and m_ℓ are given by Eqs. 10.2.24, 10.2.26, and 10.2.14, respectively. Furthermore, L is an associated Laguerre polynomial (Table 10.2) and P is an associated Legendre function (Table 10.1).

- Unfortunately, the eigenstates are only the stationary states at $t = 0$. In general, stationary states are given by Eq. 9.3.13, so

$$\psi_{n\ell}^{m_\ell} = \Psi_{n\ell}^{m_\ell}\,e^{iZ^2(13.6\text{ eV})t/\left(\hbar\,n^2\right)} \quad (10.2.29)$$

where m_ℓ is acting act as a label (*not* an exponent) on the variable $\psi_{n\ell}^{m_\ell}$. This is similar to contravariant indices from Chapter 6, but raising and lowering is irrelevant because $\psi_{n\ell}^{m_\ell}$ is *not* a rank-3 tensor. The possible values for n, ℓ, and m_ℓ are given by Eqs. 10.2.24, 10.2.26, and 10.2.14, respectively.

Shells and Orbitals

We're aware now that electrons don't really "orbit" a nucleus, as was suggested in the Bohr model (see Section 9.1). However, the stationary states, $\psi_{n\ell}^{m_\ell}$ (Eq. 10.2.29), in an atom are often referred to as **orbitals** because people are stubborn. The labels are called quantum numbers and they each have names:

- n is the principle quantum number.
  - This determines the energy level (Eq. 10.2.25):

$$\mathcal{H}\Psi = E_n\Psi = \frac{Z^2}{n^2}\,(-13.6\text{ eV})\,\psi$$

  - All states with the same n have the *roughly* same energy.
  - We sometimes call these **shells**.
- ℓ is the azimuthal quantum number.
  - This determines the magnitude of the angular momentum:

$$L^2\Psi = \ell\,(\ell+1)\,\hbar^2\,\Psi, \tag{10.2.30}$$

    where L^2 is a quantum operator (observable).
  - We also call this an **orbital type** or a **subshell**:
    * sharp or 's' ($\ell = 0$),
    * principal or 'p' ($\ell = 1$),
    * diffuse or 'd' ($\ell = 2$),
    * fundamental or 'f' ($\ell = 3$), etc.
- m_ℓ is the magnetic quantum number.
  - This determines the orientation of the angular momentum relative to an arbitrary z-axis:

$$L_z\Psi = m_\ell\hbar\,\Psi, \tag{10.2.31}$$

    where L_z is a quantum operator (observable).
  - Sometimes we call this its **magnetic moment**, hence the m.

See Table 10.3 for some examples.

Did you notice that a prediction of the value of L^2 or L_z doesn't change the state? This shouldn't be too much of a surprise since we already know $\mathcal{H}$, L^2, L_z all commute (see Eqs. 9.3.32, 9.3.33 and 9.3.34). As a result, they all have the same set of eigenstates, which means we can measure them all at the same time. Eqs. 10.2.25, 10.2.30, and 10.2.31 give us each of their eigenvalues.

Table 10.3: Based on Eqs. 10.2.24, 10.2.26, and 10.2.14 in Example 10.2.1, these are the possible values for the three quantum numbers n, ℓ, and m_ℓ in the first four electron shells. The orbital type is also given along with the number of states for each type.

n	ℓ	Orbital Type	m_ℓ	Number of States
1	0	s	0,	1
2	0	s	0,	1
2	1	p	-1, 0, 1	3
3	0	s	0,	1
3	1	p	-1, 0, 1	3
3	2	d	-2, -1, 0, 1, 2	5
4	0	s	0,	1
4	1	p	-1, 0, 1	3
4	2	d	-2, -1, 0, 1, 2	5
4	3	f	-3, -2, -1, 0, 1, 2, 3	7

We can also say a few more things about the separated parts of the eigenstates: $R_{n\ell}$ (Eq. 10.2.22), $\Theta_\ell^{m_\ell}$ (Eq. 10.2.13), and Φ^{m_ℓ} (Eq. 10.2.9).

The radial part, $R_{n\ell}$, determines scale. In Figure 10.3, you can find graphs of the radial probability densities, R^2r^2 (the integrand of Eq. 10.2.4a), for the first four s-orbitals in the hydrogen atom. You can see an $n = 1$ electron is dramatically more likely to be found around one Bohr radius, a_0 (Eq. 10.2.16), from the nucleus than anywhere else. However, this consistency with the Bohr model quickly deteriorates since the highest peaks don't line up with Eq. 9.1.2. Figure 10.4 shows the same for the p-orbitals in the hydrogen atom. The radial equations used in Figures 10.3 and 10.4 can be found in Table 10.4.

The angular parts, $\Theta_\ell^{m_\ell}$ (Eq. 10.2.13), and Φ^{m_ℓ} (Eq. 10.2.9), tell you something about the shape of the orbital. If we combine them, then

$$\boxed{Y_\ell^{m_\ell} = \Theta_\ell^{m_\ell}\Phi^{m_\ell} = \sqrt{\frac{(2\ell+1)}{4\pi}\frac{(\ell-m_\ell)!}{(\ell+m_\ell)!}}\,[P_\ell^{m_\ell}(\cos\theta)]\,e^{im_\ell\phi}}, \quad (10.2.32)$$

where $P_\ell^{m_\ell}$ is a Legendre function (see Table 10.1). This $Y_\ell^{m_\ell}$ is referred to as a **spherical harmonic** and several example can be found in Figure 10.5. It should be noted here that there is no Z dependence. The number of protons in the nucleus has no effect on the shape of these orbitals, only their scale

($R_{n\ell}$). You can actually see what they look like if you graph $\sqrt{Y^*Y}$, which I've done for you in Figure 10.5.

If you've taken any classes covering orbitals or have looked any of this up, then the shapes in Figure 10.5 probably look a little strange to you. That's because we tend not to use spherical harmonics as a standard. Atoms are often connected to others in some kind of crystal lattice, so there tends to be a convenient set of Cartesian axes we can choose. This allows us to switch to **cubic harmonics**, which are much easier to work with because they're entirely *real*.

Cubic harmonics can be found by taking linear combinations of spherical harmonics (of the same ℓ) that eliminate the imaginary parts. For example, the cubic p-orbital ($\ell = 1$) along the x-axis is

$$\mathrm{p}_x = \frac{1}{\sqrt{2}}\left(Y_1^{-1} - Y_1^1\right) = \frac{1}{\sqrt{2}}\sqrt{\frac{3}{8\pi}}\sin\theta\ \left(e^{-i\phi} + e^{i\phi}\right)$$

Using Euler's formula (Eq. 10.2.8),

$$\begin{aligned}\mathrm{p}_x &= \frac{1}{\sqrt{2}}\sqrt{\frac{3}{8\pi}}\sin\theta\ (\cos\phi - i\sin\phi + \cos\phi + i\sin\phi)\\ &= \frac{1}{\sqrt{2}}\sqrt{\frac{3}{8\pi}}\,(2\cos\phi) = \sqrt{\frac{3}{4\pi}}\sin\theta\ \cos\phi\end{aligned}$$

and using some coordinate transformations (Eq. 1.3.1), we get

$$\mathrm{p}_x = \sqrt{\frac{3}{4\pi}}\frac{x}{\sqrt{x^2+y^2+z^2}}.$$

Be *very* careful with your negative signs in this process. It's easy to forget the extra negative you have for odd m_ℓ values. The cubic harmonics for the first three orbital types (s, p, and d) are shown in Figure 10.6.

A couple of the d-orbitals given in Figure 10.6 are labeled very strangely because we're choosing to be as descriptive as possible. The labels tell you something about what the numerator looks like in Cartesian variables as well as the orbital's orientation:

- d_{xz} is in the xz-plane,
- d_{yz} is in the yz-plane,

- $d_{x^2-y^2}$ is in the xy-plane,
- d_{xy} is in the xy-plane, and
- d_{z^2} orbital is along the z-axis.

This is just like the labels on the p-orbitals:

- p_x is along the x-axis,
- p_y is along the y-axis, and
- p_z is along the z-axis.

It's important to know what these look like because, as it turns out, they're the same shape in multiple-electron atoms.

Table 10.4: This is the first ten radial equations, $R_{n\ell}(r)$, for the hydrogen atom ($Z = 1$). They were found using Eqs. 10.2.22 and 10.2.23. These were used in Figures 10.3 and 10.4 by computing $\left(R_{n\ell}\, a^{3/2}\right)^2 r^2$.

$R_{n\ell}$		Radial Equations
R_{10}	$=$	$a_0^{-3/2}\, 2\, e^{-r/a_0}$
R_{20}	$=$	$a_0^{-3/2}\, \frac{\sqrt{2}}{4} \left[2 - \left(\frac{r}{a_0}\right)\right] e^{-r/(2a_0)}$
R_{21}	$=$	$a_0^{-3/2}\, \frac{\sqrt{6}}{12} \left(\frac{r}{a_0}\right) e^{-r/(2a_0)}$
R_{30}	$=$	$a_0^{-3/2}\, \frac{2\sqrt{3}}{27} \left[3 - 2\left(\frac{r}{a_0}\right) + \frac{2}{9}\left(\frac{r}{a_0}\right)^2\right] e^{-r/(3a_0)}$
R_{31}	$=$	$a_0^{-3/2}\, \frac{\sqrt{6}}{81} \left(\frac{r}{a_0}\right) \left[4 - \frac{2}{3}\left(\frac{r}{a_0}\right)\right] e^{-r/(3a_0)}$
R_{32}	$=$	$a_0^{-3/2}\, \frac{2\sqrt{30}}{1{,}215} \left(\frac{r}{a_0}\right)^2 e^{-r/(3a_0)}$
R_{40}	$=$	$a_0^{-3/2}\, \frac{1}{16} \left[4 - 3\left(\frac{r}{a_0}\right) + \frac{1}{2}\left(\frac{r}{a_0}\right)^2 - \frac{1}{48}\left(\frac{r}{a_0}\right)^3\right] e^{-r/(4a_0)}$
R_{41}	$=$	$a_0^{-3/2}\, \frac{\sqrt{15}}{480} \left(\frac{r}{a_0}\right) \left[10 - \frac{5}{2}\left(\frac{r}{a_0}\right) + \frac{1}{8}\left(\frac{r}{a_0}\right)^2\right] e^{-r/(4a_0)}$
R_{42}	$=$	$a_0^{-3/2}\, \frac{\sqrt{5}}{1{,}920} \left(\frac{r}{a_0}\right)^2 \left[6 - \frac{1}{2}\left(\frac{r}{a_0}\right)\right] e^{-r/(4a_0)}$
R_{43}	$=$	$a_0^{-3/2}\, \frac{\sqrt{35}}{26{,}880} \left(\frac{r}{a_0}\right)^3 e^{-r/(4a_0)}$

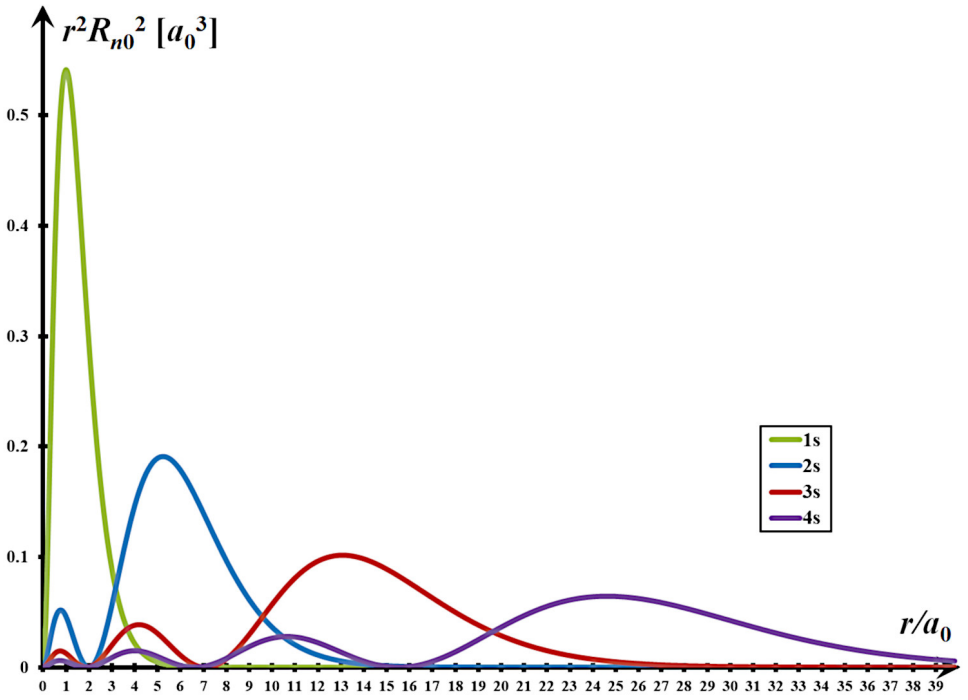


Figure 10.3: This graph shows the probability densities of the first four $R(r)$ (Eq. 10.2.22) functions for $\ell = 0$ (i.e. the s-orbitals) in the hydrogen atom.

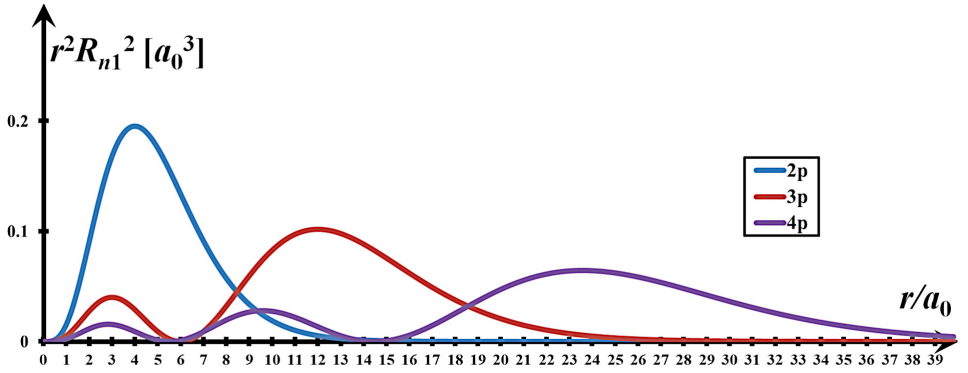


Figure 10.4: This graph shows the probability densities of the first three $R(r)$ (Eq. 10.2.22) functions for $\ell = 1$ (i.e. the p-orbitals) in the hydrogen atom. Note: The $n = 1$ energy level does *not* have a p-orbital.

$Y_\ell^{m_\ell}$	Spherical Harmonics		
Y_0^0	$= \sqrt{\frac{1}{4\pi}}$		
Y_1^0	$= \sqrt{\frac{3}{4\pi}}\cos\theta$	$Y_1^{\pm1}$	$= \mp\sqrt{\frac{3}{8\pi}}\sin\theta\, e^{\pm i\phi}$
Y_2^0	$= \sqrt{\frac{5}{16\pi}}\left(3\cos^2\theta - 1\right)$	$Y_2^{\pm1}$	$= \mp\sqrt{\frac{15}{8\pi}}\cos\theta\,\sin\theta\, e^{\pm i\phi}$
$Y_2^{\pm2}$	$= \sqrt{\frac{15}{32\pi}}\sin^2\theta\, e^{\pm 2i\phi}$		
Y_3^0	$= \sqrt{\frac{7}{16\pi}}\left(5\cos^3\theta - 3\cos\theta\right)$	$Y_3^{\pm1}$	$= \mp\sqrt{\frac{21}{64\pi}}\left(5\cos^2\theta - 1\right)\sin\theta\, e^{\pm i\phi}$
$Y_3^{\pm2}$	$= \sqrt{\frac{105}{32\pi}}\cos\theta\,\sin^2\theta\, e^{\pm 2i\phi}$	$Y_3^{\pm3}$	$= \mp\sqrt{\frac{35}{64\pi}}\sin^3\theta\, e^{\pm 3i\phi}$

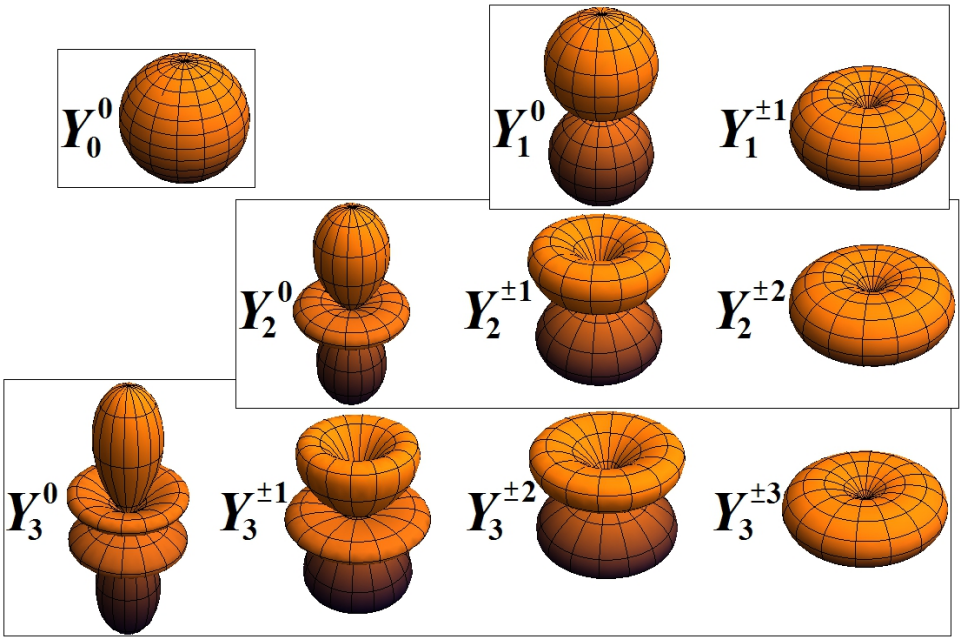


Figure 10.5: This is a visual representation of $\sqrt{Y^*Y}$ for the spherical harmonics (Eq. 10.2.32). Only those for the first 4 values of ℓ are shown. Note: $Y_\ell^{m_\ell}$ looks the same as $Y_\ell^{-m_\ell}$ because all negatives disappear in the complex square.

Orbital		Cubic Harmonics		
s	$=$	Y_0^0	$=$	$\sqrt{\frac{1}{4\pi}}$
p_x	$=$	$\frac{1}{\sqrt{2}}\left(Y_1^{-1}-Y_1^1\right)$	$=$	$\sqrt{\frac{3}{4\pi}}\frac{x}{\sqrt{x^2+y^2+z^2}}$
p_y	$=$	$i\frac{1}{\sqrt{2}}\left(Y_1^{-1}+Y_1^1\right)$	$=$	$\sqrt{\frac{3}{4\pi}}\frac{y}{\sqrt{x^2+y^2+z^2}}$
p_z	$=$	Y_1^0	$=$	$\sqrt{\frac{3}{4\pi}}\frac{z}{\sqrt{x^2+y^2+z^2}}$
d_{xz}	$=$	$\frac{1}{\sqrt{2}}\left(Y_2^{-1}-Y_2^1\right)$	$=$	$\sqrt{\frac{15}{4\pi}}\frac{xz}{x^2+y^2+z^2}$
d_{yz}	$=$	$i\frac{1}{\sqrt{2}}\left(Y_2^{-1}+Y_2^1\right)$	$=$	$\sqrt{\frac{15}{4\pi}}\frac{yz}{x^2+y^2+z^2}$
d_{xy}	$=$	$i\frac{1}{\sqrt{2}}\left(Y_2^{-2}-Y_2^2\right)$	$=$	$\sqrt{\frac{15}{16\pi}}\frac{xy}{x^2+y^2+z^2}$
$\mathrm{d}_{x^2-y^2}$	$=$	$\frac{1}{\sqrt{2}}\left(Y_2^{-2}+Y_2^2\right)$	$=$	$\sqrt{\frac{15}{16\pi}}\frac{x^2-y^2}{x^2+y^2+z^2}$
d_{z^2}	$=$	Y_2^0	$=$	$\sqrt{\frac{5}{16\pi}}\left(\frac{3z^2}{x^2+y^2+z^2}-1\right)$

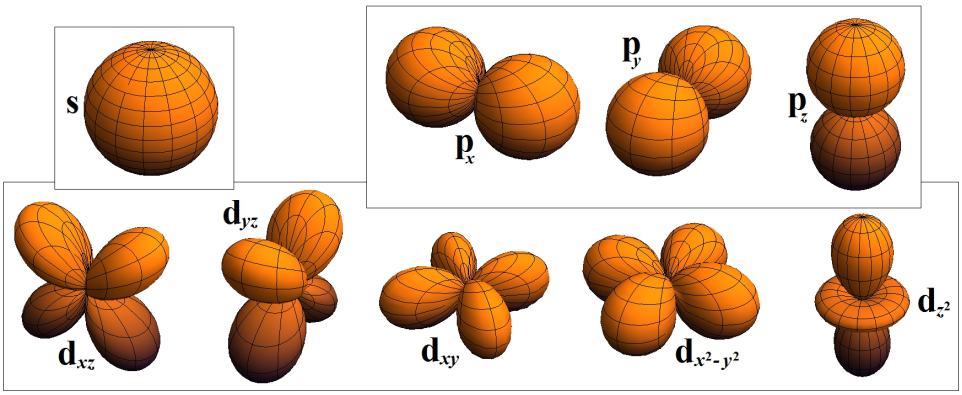


Figure 10.6: This is the first nine cubic harmonics where $Y_\ell^{m_\ell}$ is a spherical harmonic from Figure 10.5. All orbitals for the first three types (s, p, and d) are shown. The transformations in Eq. 1.3.1 were used to get functions of x, y, and z. Each of them is named for the Cartesian numerator.

Spin Angular Momentum

We saw in Eqs. 10.2.30 and 10.2.31 that the electron has an angular momentum, often called **orbital** angular momentum because it's related to the orbital type (i.e. the value of ℓ). It's a property the electron has because of it's *behavior*, so we call it an **extrinsic property**. Many people studying quantum mechanics imagine it's like the Earth orbiting the Sun, but this is mistake. The electron doesn't really "orbit" the nucleus. It simply exists in an "orbital." Don't make the analogy just because the names look the same.

Another property electrons have is spin angular momentum or just **spin**. I acknowledged it's existence a few times in Chapter 9, but we haven't been ready to discuss it until now. Again, do *not* make the analogy with the Earth! The electron is not "spinning." We only call this "spin" because we're used to hearing words like that when dealing with angular momentum. The electron has this spin *regardless of it's behavior*, so we call it an **intrinsic property**. Like charge, it just *has* it.

Spin is something we can measure for all particles. Mathematically, it behaves a lot like orbital angular momentum. Recalling Eqs. 9.3.34, 9.3.38, 10.2.30, and 10.2.31;

- Commutator between Spin and Spin along z:

$$\left[S^2, S_z\right] = 0 \tag{10.2.33}$$

  where S^2 and S_z are both quantum operators (observables).

- Commutator between components of Spin:

$$[S_i, S_j] = i\hbar\varepsilon_{ijk}S_k \tag{10.2.34}$$

  where S_i and S_j are both quantum operators (observables) and ε_{ijk} is the Levi-Civita pseudotensor (Eq. 6.6.4).

- Prediction of the magnitude of Spin:

$$S^2\left|s, m_s\right\rangle = s\left(s+1\right)\hbar^2\left|s, m_s\right\rangle, \tag{10.2.35}$$

  where S^2 is a quantum operator (observable).

- Prediction of the orientation of Spin relative to an arbitrary z-axis:

$$S_z\left|s, m_s\right\rangle = m_s\hbar\left|s, m_s\right\rangle, \tag{10.2.36}$$

  where S_z is a quantum operator (observable).

We categorize particles by their spin quantum number, s, because the value never changes. Like ℓ, it does have restrictions:

$$s = 0,\ \frac{1}{2},\ 1,\ \frac{3}{2},\ 2,\ \ldots \tag{10.2.37}$$

Although, unlike ℓ, it can take half-integer values and has no other quantum number to give it an upper limit. The quantum number m_s is restricted just like m_ℓ:

$$m_s = -s,\ -s+1,\ \ldots,\ s-1,\ s; \tag{10.2.38}$$

taking on values from $-s$ to $+s$ in increments of one. However, for *massless* particles, m_s can only take on the extreme values $-s$ and $+s$ (e.g. m_s for a photon is either -1 or 1, but not zero).

The state function, ψ, has been replaced with a ket vector, $|s, m_s\rangle$, for convenience. As mentioned in Example 9.4.5, since spin is discrete (rather than continuous), it makes more sense to use bra-ket notation (rather than function/integral notation). For electrons and protons, $s = 1/2$, so we call them **spin-$\frac{1}{2}$ particles**. That means m_s can have only two values, $\pm 1/2$, and the only available states are

$$\left|\tfrac{1}{2}, +\tfrac{1}{2}\right\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \left|\tfrac{1}{2}, -\tfrac{1}{2}\right\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{10.2.39}$$

or "spin-up" ($m_s = +1/2$) and "spin-down" ($m_s = -1/2$).

If we're using vectors for the spin states (called **spinors**), then it's also convenient to write the quantum operators as matrices:

$$S_x = \frac{\hbar}{2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\ S_y = \frac{\hbar}{2}\begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix},\ \text{and } S_z = \frac{\hbar}{2}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \tag{10.2.40}$$

where $S^2 \equiv S_x^2 + S_y^2 + S_z^2$. Note: Eqs. 10.2.39 and 10.2.40 are completely consistent with Eqs. 10.2.35 and 10.2.36. These matrices are always square and have a number of rows (and columns) equal to the number of possible values for m_s. I'll save a discussion of other values of spin for Appendix D.

Full Angular Momentum

If circumstances require you consider effects involving both orbital *and* spin angular momentum, then problems ensue. Under these considerations, $\mathcal{H}$

still commutes with L^2 or S^2, so Eqs. 10.2.30 and 10.2.35 are still valid. However, $\mathcal{H}$ no longer commutes with L_z or S_z, making the use of quantum numbers m_ℓ and m_s undesirable. It also invalidates Eq. 10.2.31 because L_z and $\mathcal{H}$ no longer share the same eigenstates, Ψ (Eq. 10.2.27). Basically, we can't make predictions about the orientation of $\vec{L}$ or $\vec{S}$ at the same time we make predictions about the energy.

Fortunately, we can solve this problem by adding them together. We'll define a **full angular momentum**,

$$\vec{J} \equiv \vec{S} + \vec{L}, \tag{10.2.41}$$

with a magnitude measured by J^2 and orientation measured by J_z, both of which commute with $\mathcal{H}$. Mathematically, the full angular momentum behaves just like orbital or spin angular momentum.

- Commutator between the magnitude and orientation:

$$\left[J^2, J_z\right] = 0 \tag{10.2.42}$$

  where J^2 and J_z are both quantum operators (observables).

- Commutator between components:

$$[J_i, J_j] = i\hbar\varepsilon_{ijk}J_k \tag{10.2.43}$$

  where J_i and J_j are both quantum operators (observables) and ε_{ijk} is the Levi-Civita pseudotensor (Eq. 6.6.4).

- Prediction of the magnitude:

$$J^2 \left|j, m_j\right\rangle = j\left(j+1\right)\hbar^2 \left|j, m_j\right\rangle, \tag{10.2.44}$$

  where J^2 is a quantum operator (observable).

- Prediction of the orientation:

$$J_z \left|j, m_j\right\rangle = m_j\hbar \left|j, m_j\right\rangle, \tag{10.2.45}$$

  where J_z is a quantum operator (observable).

The magnitude quantum number j can take on the following values:

$$j = |\ell - s|\,,\ |\ell - s| + 1,\ \ldots,\ (\ell + s) - 1,\ (\ell + s)\,; \tag{10.2.46}$$

or the values between $|\ell - s|$ and $(\ell + s)$ in increments of one. For an electron, the only possible values are $j = |\ell \pm 1/2|$. The orientation quantum number m_j is restricted just like m_s and m_ℓ:

$$m_j = -j,\ -j + 1,\ \ldots,\ j - 1,\ j; \tag{10.2.47}$$

taking on values from $-j$ to $+j$ in increments of one.

The full angular momentum states in Eqs. 10.2.44 and 10.2.45 are shown as $|j, m_j\rangle$, which is very similar to the spin states: $|s, m_s\rangle$. We could have even written the orbital angular momentum states as $|\ell, m_\ell\rangle$, such that

$$\langle\theta, \phi|\, |\ell, m_\ell\rangle = Y_\ell^{m_\ell}, \tag{10.2.48}$$

where $Y_\ell^{m_\ell}$ are the spherical harmonics (Eq. 10.2.32). The spherical harmonics are still eigenstates of L_z! That means you could write Eq. 10.2.30 and 10.2.31 as

$$\begin{array}{rcl} L^2\, Y_\ell^{m_\ell} & = & \ell\,(\ell + 1)\,\hbar^2\, Y_\ell^{m_\ell} \\ & \text{or} & \\ L^2\, |\ell, m_\ell\rangle & = & \ell\,(\ell + 1)\,\hbar^2\, |\ell, m_\ell\rangle \end{array} \tag{10.2.49}$$

and

$$\begin{array}{rcl} L_z\, Y_\ell^{m_\ell} & = & m_\ell \hbar\, Y_\ell^{m_\ell} \\ & \text{or} & \\ L_z\, |\ell, m_\ell\rangle & = & m_\ell \hbar\, |\ell, m_\ell\rangle \end{array}, \tag{10.2.50}$$

which are always true. Since L_z and S_z commute with each other, m_ℓ and m_s can be predicted at the same time. However, since neither commutes with $\mathcal{H}$, there's no guarantee either can be predicted at the same time as n, ℓ, s, j, or m_j (which can all be predicted together). The consequence is that $|j, m_j\rangle$ is an eigenstate of the Hamiltonian, but $|\ell, m_\ell\rangle\, |s, m_s\rangle$ is not likely to be.

If the particle is in an eigenstate of the Hamiltonian, then m_ℓ and m_s are *not* definite, so $|j, m_j\rangle$ must be some linear combination:

$$|j, m_j\rangle = \sum_{m_\ell + m_s = m_j} C^{\ell,s,j}_{m_\ell, m_s, m_j}\, |\ell, m_\ell\rangle\, |s, m_s\rangle\,. \tag{10.2.51}$$

Table 10.5: This is a small sample of the Clebsch-Gordan coefficients, $C^{\ell,s,j}_{m_\ell,m_s,m_j}$, corresponding to a spin-$\frac{1}{2}$ particle in a p-orbital ($\ell = 1$). Remember, $m_j = m_\ell + m_s$ or the coefficient is zero.

$\ell = 1$	$s = \frac{1}{2}$						
		$\lvert j, m_j\rangle$					
$\lvert \ell, m_\ell\rangle$	$\lvert s, m_s\rangle$	$\lvert\frac{3}{2},+\frac{3}{2}\rangle$	$\lvert\frac{3}{2},+\frac{1}{2}\rangle$	$\lvert\frac{1}{2},+\frac{1}{2}\rangle$	$\lvert\frac{3}{2},-\frac{1}{2}\rangle$	$\lvert\frac{1}{2},-\frac{1}{2}\rangle$	$\lvert\frac{3}{2},-\frac{3}{2}\rangle$
$\lvert 1,+1\rangle$	$\lvert\frac{1}{2},+\frac{1}{2}\rangle$	1	0	0	0	0	0
$\lvert 1,+1\rangle$	$\lvert\frac{1}{2},-\frac{1}{2}\rangle$	0	$\sqrt{\frac{1}{3}}$	$\sqrt{\frac{2}{3}}$	0	0	0
$\lvert 1,0\rangle$	$\lvert\frac{1}{2},+\frac{1}{2}\rangle$	0	$\sqrt{\frac{2}{3}}$	$-\sqrt{\frac{1}{3}}$	0	0	0
$\lvert 1,0\rangle$	$\lvert\frac{1}{2},-\frac{1}{2}\rangle$	0	0	0	$\sqrt{\frac{2}{3}}$	$\sqrt{\frac{1}{3}}$	0
$\lvert 1,-1\rangle$	$\lvert\frac{1}{2},+\frac{1}{2}\rangle$	0	0	0	$\sqrt{\frac{1}{3}}$	$-\sqrt{\frac{2}{3}}$	0
$\lvert 1,-1\rangle$	$\lvert\frac{1}{2},-\frac{1}{2}\rangle$	0	0	0	0	0	1

The sum is taken over all values of m_ℓ and m_s such that $m_\ell + m_s = m_j$ (required by Eq. 10.2.41). The coefficients, C, are called **Clebsch-Gordan coefficients** and the explicit formula is horrendous, so it's usually best to look them up (see Table 10.5). This expansion process also works in reverse. If you already measured both m_ℓ and m_s, then you'll know m_j for certain, but not j. This means $\lvert \ell, m_\ell\rangle \lvert s, m_s\rangle$ must be some linear combination:

$$\lvert \ell, m_\ell\rangle \lvert s, m_s\rangle = \sum_j C^{\ell,s,j}_{m_\ell,m_s,m_j} \lvert j, m_j\rangle , \tag{10.2.52}$$

which is helpful if you want to operate on $\lvert \ell, m_\ell\rangle \lvert s, m_s\rangle$ with J_z.

Example 10.2.2

An electron is in a p-orbital for which you've already measured the full angular momentum ($j = 1/2$ and $m_j = -1/2$). Expand this state, find the probabilities for each value of m_ℓ, and calculate $\langle L_z\rangle$.

- It's an electron, so $s = 1/2$. This means we have two possibles for m_s: $\pm 1/2$.

- It's in a p-orbital, so $\ell = 1$. However, we don't know which one. All we know is $m_j = -1/2$, so we can could have either $m_\ell = 0$ or $m_\ell = -1$, one for each value of m_s such that $m_\ell = m_j - m_s$.

- Using Eq. 10.2.51 and Table 10.5, we get

$$\begin{aligned}\left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle &= \sum_{m_\ell+m_s=-\frac{1}{2}} C^{1,\frac{1}{2},\frac{1}{2}}_{m_\ell,m_s,-\frac{1}{2}} \left|1,m_\ell\right\rangle \left|\tfrac{1}{2},m_s\right\rangle \\ &= \sqrt{\frac{1}{3}}\left|1,0\right\rangle\left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle - \sqrt{\frac{2}{3}}\left|1,-1\right\rangle\left|\tfrac{1}{2},+\tfrac{1}{2}\right\rangle.\end{aligned}$$

- Since this is bra-ket notation, we use Eq. 9.3.8 to find probability. The probability of finding the electron in $m_\ell = 0$ is

$$\begin{aligned}P &= \left\|\langle \ell, m_\ell|\langle s, m_s|\, |j, m_j\rangle\right\|^2 = \left\|\langle 1,0|\left\langle\tfrac{1}{2},-\tfrac{1}{2}\right|\left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle\right\|^2 \\ &= \left\|\langle 1,0|\left\langle\tfrac{1}{2},-\tfrac{1}{2}\right|\left(\sqrt{\frac{1}{3}}\left|1,0\right\rangle\left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle - \sqrt{\frac{2}{3}}\left|1,-1\right\rangle\left|\tfrac{1}{2},+\tfrac{1}{2}\right\rangle\right)\right\|^2 \\ &= \left\|\sqrt{\frac{1}{3}}(1) - \sqrt{\frac{2}{3}}(0)\right\|^2 = \frac{1}{3}.\end{aligned}$$

  By similar work, the probability of finding the electron in $m_\ell = -1$ is $2/3$. This makes sense because it should be in *one* of them and $1 - 1/3 = 2/3$. We'll save any further interpretation for Section 10.4.

- The expectation value of L_z is given by

$$\langle L_z\rangle = \langle j, m_j|\, L_z\, |j, m_j\rangle = \left\langle\tfrac{1}{2},-\tfrac{1}{2}\right|\, L_z\, \left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle,$$

  but we need to expand into $|\ell, m_\ell\rangle$. First, we'll operate L_z on the ket vector:

$$\begin{aligned}L_z\left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle &= L_z\left(\sqrt{\frac{1}{3}}\left|1,0\right\rangle\left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle - \sqrt{\frac{2}{3}}\left|1,-1\right\rangle\left|\tfrac{1}{2},+\tfrac{1}{2}\right\rangle\right) \\ &= \sqrt{\frac{1}{3}}(0)\left|1,0\right\rangle\left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle - \sqrt{\frac{2}{3}}(-\hbar)\left|1,-1\right\rangle\left|\tfrac{1}{2},+\tfrac{1}{2}\right\rangle \\ &= \hbar\sqrt{\frac{2}{3}}\left|1,-1\right\rangle\left|\tfrac{1}{2},+\tfrac{1}{2}\right\rangle\end{aligned}$$

where we've used Eq. 10.2.50 to operate. Operating with the bra vector,

$$\langle j, m_j| = \left\langle \tfrac{1}{2}, -\tfrac{1}{2}\right| = \sqrt{\frac{1}{3}}\,\langle 1, 0|\left\langle \tfrac{1}{2}, -\tfrac{1}{2}\right| - \sqrt{\frac{2}{3}}\,\langle 1, -1|\left\langle \tfrac{1}{2}, +\tfrac{1}{2}\right|,$$

we get

$$\langle L_z\rangle = \sqrt{\frac{1}{3}}\,\hbar\sqrt{\frac{2}{3}}\,(0) - \sqrt{\frac{2}{3}}\,\hbar\sqrt{\frac{2}{3}}\,(1) = -\frac{2\hbar}{3}.$$

- Remember, this is an average of all possible values of L_z *weighted* by the probabilities of each. We could have just easily said

$$\langle L_z\rangle = \frac{1}{3}\,(0) + \frac{2}{3}\,(-\hbar) = -\frac{2\hbar}{3},$$

which is the same result.

Example 10.2.3

An electron is in a p_z-orbital for which you've already measured it to be spin-down. Expand this state into $|j, m_j\rangle$.

- It's an electron, so $s = 1/2$. It's also spin-down, so $m_s = -1/2$.
- It's in a p_z-orbital, so we know $\ell = 1$ and $m_\ell = 0$.
- Since $m_j = m_\ell + m_s$, we know $m_j = -1/2$ is the only option. However, j is not definite. By Eq. 10.2.46, the available options are $j = 1/2$ and $j = 3/2$.
- Using Eq. 10.2.52 and Table 10.5, we get

$$\begin{aligned} |1, 0\rangle\left|\tfrac{1}{2}, -\tfrac{1}{2}\right\rangle &= \sum_j C^{1,\frac{1}{2},j}_{0,-\frac{1}{2},m_j}\ |j, m_j\rangle \\ &= \sqrt{\frac{2}{3}}\ \left|\tfrac{3}{2}, -\tfrac{1}{2}\right\rangle + \sqrt{\frac{1}{3}}\ \left|\tfrac{1}{2}, -\tfrac{1}{2}\right\rangle. \end{aligned}$$

Fine Structure

We mentioned several times the energy of an electron in shell n has an energy given by Eq. 10.2.25. This implies that the electron can have *any* possible value for ℓ or m_ℓ for that n and have *exactly* the same energy. When more than one stationary state has the same energy, we say the model has **degeneracy**. The three-dimensional infinite well (Eq. 9.4.15) and the three-dimensional harmonic oscillator (Eq. 9.4.87) had this same problem, so it might seem commonplace when working in three dimensions.

This isn't really true though. In deriving the stationary states for single-electron atoms in Example 10.2.1, we unwittingly made some assumptions and we all know what happens when we assume. Here is a list of those assumptions:

1. the nucleus was stationary,
2. the electron was non-relativistic,
3. the electron had no spin,
4. the proton had no spin, and
5. the Coulomb potential energy (Eq. 10.2.1) was continuous.

These were great *approximations* for getting us simple stationary states like those in Eq. 10.2.29, but we need to be careful about the conclusions we take away from approximate results.

Now, if we hadn't made assumptions 2-5, then Schrödinger's equation (Eq. 9.2.7) would have been analytically unsolvable. I'm not suggesting we start over and do this numerically (although, you could). I'm just saying we can get *closer* to reality by adjusting our results a bit. First, we'll define the **fine structure constant**, which is

$$\alpha = \frac{q^2}{4\pi\hbar c\epsilon_0} = \frac{\hbar}{a_0 mc} = 7.29735257 \times 10^{-3} = \frac{1}{137.036}, \tag{10.2.53}$$

according to 2014 CODATA recommended values. It's a unitless quantity, so named because it's involved in very small adjustments to the energy levels. Using this for the hydrogen atom ($Z = 1$), the energy levels (Eq. 10.2.25) can be written as

$$E_n = -\frac{1}{n^2}\frac{\hbar^2}{2ma_0^2} = -\frac{\alpha^2 mc^2}{2n^2}, \tag{10.2.54}$$

where $m = m_e$ is the mass of the electron and c is the speed of light (Eq. 5.5.4). This gives us a basis for comparison.

We're going to keep things as straightforward as possible by handling one approximation at a time. The following list is in the same order as the list above and applies only to hydrogen ($Z = 1$):

1. *The nucleus is not stationary.* It does wiggle a little in response to the tug of the electron. In the hydrogen atom, $m_{\text{nuc}} = 1836 m_e$, so its response (i.e. its acceleration) is 1836 times smaller because of Newton's second law (Eq. 4.2.6). This factor is even larger when the nucleus is bigger, so it was a pretty decent assumption to make. However, if you want (or need) to be more accurate, then you just need to use the **reduced mass** for the electron:

$$m \quad \to \quad \mu = \left(\frac{m_{\text{nuc}}}{m + m_{\text{nuc}}} \right) m, \tag{10.2.55}$$

where $m = m_e$ is the mass of the electron. Since E_n is proportional to $m = m_e$, this slightly reduces the value for energy by

$$\Delta E_{n,\mu} = E_{\text{new}} - E_{\text{old}} = \left(\tfrac{m_{\text{nuc}}}{m+m_{\text{nuc}}} \right) E_n - E_n = \left(\tfrac{m_{\text{nuc}}}{m+m_{\text{nuc}}} - 1 \right) E_n$$

$$\boxed{\Delta E_{n,\mu} = \left(\tfrac{1836}{1837} - 1 \right) E_n = -5.444 \times 10^{-4} E_n}, \tag{10.2.56}$$

which is a factor of $\approx 10^{-3}$ in an **order of magnitude approximation** (see Section A.3 for more details). It might seem silly to discuss adjustments this small, but we need to if we intend on understanding what's *really* happening. Believe it or not, this is the biggest adjustment we're going to see.

2. *Technically, all particles are relativistic.* Relativity applies to all particles all the time. We just decide that when $v \ll c$, we don't need to go through the trouble. Unfortunately, the small size of the adjustments we're making to E_n require it. If we started over with Schrödinger's equation (Eq. 9.2.7), then the kinetic energy can be found using Eq. 7.4.25 and the Hamiltonian would be

$$\mathcal{H} = KE + PE = [E_{\text{rel}} - E_p] + V$$

$$\mathcal{H} = mc^2 \left[\sqrt{1 + \frac{p_{\text{rel}}^2}{m^2 c^2}} - 1 \right] + V. \tag{10.2.57}$$

Traditionally, this is solved using an approximation method called **perturbation theory** (a *very* poor use of the word "theory"). The advent of computers has made this method a bit obsolete, so I'll spare you the unnecessary pain of showing you how it works. The result is an adjustment in energy of

$$\Delta E_{n,\text{rel}} \approx \frac{\alpha^2}{2n^2} \left(\frac{4n}{2\ell + 1} - \frac{3}{2} \right) E_n, \tag{10.2.58}$$

where the factor in front depends on the state (i.e. the quantum numbers n and ℓ). However, if we do another order of magnitude approximation, we get $\approx 10^{-4}$ to 10^{-6} for the states the electron is found in the most often (i.e. the states near the nucleus).

3. *The electron is a spin-$\frac{1}{2}$ particle.* All forms of angular momentum, including spin, generate something we call a magnetic moment or **magnetic dipole moment**. According to Faraday's law (Eq. 5.3.11), a changing magnetic field produces an electric field. Since the electron is a *moving* magnetic dipole, it produces an **electric dipole moment**. The proton was already exerting an electric force on the electron due to its charge, $q = q_e$, but there is now an *additional* force due to the electric dipole moment.

   The math in the rest frame of the proton is a bit challenging, so we usually do this in the rest frame of the electron (an accelerated frame). There, the electron is stationary, so no electric dipole moment. However, the proton is now in an orbital of the electron, so it's motion produces a *magnetic* field that can influence the electrons *magnetic* dipole moment. The result is an additional term in the Hamiltonian:

$$\mathcal{H} = mc^2 \left[\sqrt{1 + \frac{p_{\text{rel}}^2}{m^2 c^2}} - 1 \right] + V + \frac{q^2}{8\pi m^2 c^2 \epsilon_0} \left(\vec{S} \bullet \vec{L} \right) \tag{10.2.59}$$

where $\vec{S} \bullet \vec{L} = S_x L_x + S_y L_y + S_z L_z$ (i.e. it's an operator). We call this **spin-orbit coupling** because we can no longer consider them

separately. We've already defined a solution to this problem: the full angular momentum, $\vec{J}$. By Eq. 10.2.41,

$$\vec{S} \bullet \vec{L} = \frac{1}{2}\left(J^2 - S^2 - L^2\right), \tag{10.2.60}$$

which allows us to avoid using operators that don't commute with $\mathcal{H}$. This additional term demands an adjustment in energy of

$$\Delta E_{n,\text{so}} \approx -\frac{\alpha^2}{2n^2}\left(\frac{2n\left[j\left(j+1\right) - \ell\left(\ell+1\right) - 3/4\right]}{\ell\left(2\ell+1\right)\left(\ell+1\right)}\right) E_n, \tag{10.2.61}$$

where the factor in front depends on the state (i.e. the quantum numbers n, ℓ, and j). Performing another order of magnitude approximation, we get $\approx 10^{-4}$ to 10^{-6} for the states the electron is found in the most often (i.e. the states near the nucleus). This is the same as the relativistic adjustment, so we'll add Eq. 10.2.58 to Eq. 10.2.61 to get

$$\boxed{\Delta E_{n,\text{fs}} \approx \frac{\alpha^2}{2n^2}\left(\frac{4n}{2j+1} - \frac{3}{2}\right) E_n}. \tag{10.2.62}$$

This is called the **fine structure adjustment** and depends only on n and j (i.e. the pure ℓ dependence in Eq. 10.2.58 has been canceled by Eq. 10.2.61).

4. *The proton is a spin-$\frac{1}{2}$ particle.* All forms of angular momentum, including spin, generate something we call a magnetic moment or **magnetic dipole moment**. If both the electron and proton have spin, then they will both have a magnetic moment causing yet *another* magnetic interaction. This demands more terms in the Hamiltonian:

$$\mathcal{H} = \mathcal{H}_{\text{fs}} + \frac{5.59\,\mu_0\,q^2}{8\pi m_p m_e}\left[\frac{3(\vec{S}_p \bullet \hat{r})(\vec{S}_e \bullet \hat{r}) - \vec{S}_p \bullet \vec{S}_e}{r^3}\right] + \frac{5.59\,\mu_0\,q^2}{3m_p m_e}\left(\vec{S}_p \bullet \vec{S}_e\right)\delta^3(\vec{r}) \tag{10.2.63}$$

where $\vec{S}_p$ is the proton spin, $\vec{S}_e$ is the electron spin, $\mathcal{H}_{\text{fs}}$ is given by Eq. 10.2.59, and $\delta^3(\vec{r})$ is the Dirac delta function (Eq. 5.3.7). This is called **spin-spin coupling** because we can no longer consider the spins separately. The full spin is defined as

$$\vec{S} \equiv \vec{S}_p + \vec{S}_e, \tag{10.2.64}$$

similar to Eq. 10.2.41. Similar to spin-orbit coupling,

$$\vec{S}_p \bullet \vec{S}_e = \frac{1}{2}\left(S^2 - S_p^2 - S_e^2\right), \tag{10.2.65}$$

which allows us to avoid using operators that don't commute with $\mathcal{H}$. A consequence is that s and m_s now describe the full spin, not the individual spins of the particles.

Since the operations themselves depend on ℓ, the solution must be piecewise:

$$\boxed{\Delta E_{n,\text{hf}} \approx -\frac{11.18\alpha^2}{n}\frac{m_e}{m_p}E_n \begin{cases} \frac{1}{6}(2j\pm 1)(2j\pm 1+2) - 1 & \text{, if } \ell = 0 \\ \dfrac{\pm 1}{(2j\pm 1+1)(2\ell+1)} & \text{, if } \ell \neq 0 \end{cases}} \tag{10.2.66}$$

which is dependent on n, ℓ, and j. Performing another order of magnitude approximation, we get $\approx 10^{-6}$ to 10^{-7} for the s-orbitals ($\ell = 0$) near the nucleus and $\approx 10^{-8}$ to 10^{-9} for $\ell \neq 0$ near the nucleus. That's a few orders smaller than the fine structure adjustment, so we call this the **hyperfine structure adjustment**.

5. *The Coulomb potential energy is discrete like the energy of the electron.* If we revisit Eq. 10.2.1, we can see that $V(r) \to -\infty$ as $r \to 0$. We know how to deal with infinities mathematically, so this wasn't a problem in getting basic results. However, observation has shown us that nothing in the universe is *really* infinite, so V must have a minimum value. This is accomplished by realizing V is **quantized** (i.e. it takes on only discrete values).

   From Table 10.4, we can see $R_{n\ell}(0) = 0$ for any $\ell > 0$ (i.e. the electron is *never* there), so this adjustment is much larger for s-orbitals ($\ell = 0$). Even so, we can find a general solution for all orbitals experimentally. It's called the **Lamb shift** and is given by

$$\boxed{\Delta E_{n,\text{lamb}} \approx -\frac{\alpha^3}{2n}E_n \begin{cases} 13 & \text{, if } \ell = 0 \\ 0.05 \pm \dfrac{4}{\pi(2j+1)(2\ell+1)} & \text{, if } \ell \neq 0 \end{cases}} \tag{10.2.67}$$

which is dependent on n, ℓ, and j. For the $\pm$ in the $\ell \neq 0$ case, plus is for $j = \ell + 1/2$ and minus is for $j = \ell - 1/2$. Performing another order

Table 10.6: This is a summary of the (fine and hyperfine) adjustments to the energy levels, E_n (Eq. 10.2.25), in the hydrogen atom. Each is given as an order of magnitude for simplicity and clarity.

Category	Description		Order of Magnitude
	Reduced Mass	$\approx$	$10^{-3}E_n$
Fine Structure	Relativistic	$\approx$	$10^{-4}E_n$ to $10^{-6}E_n$
	Spin-Orbit Coupling	$\approx$	$10^{-4}E_n$ to $10^{-6}E_n$
Hyperfine Structure	Spin-Spin Coupling		
	for s-orbitals ($\ell = 0$)	$\approx$	$10^{-6}E_n$ to $10^{-7}E_n$
	for p, d, f, etc. ($\ell > 0$)	$\approx$	$10^{-8}E_n$ to $10^{-9}E_n$
Lamb Shift	Quantized V		
	for s-orbitals ($\ell = 0$)	$\approx$	$10^{-6}E_n$ to $10^{-7}E_n$
	for p, d, f, etc. ($\ell > 0$)	$\approx$	$10^{-8}E_n$ to $10^{-9}E_n$

of magnitude approximation, we get $\approx 10^{-6}$ to 10^{-7} for the s-orbitals ($\ell = 0$) near the nucleus and $\approx 10^{-8}$ to 10^{-9} for $\ell \neq 0$ near the nucleus. This is about the same as the hyperfine adjustment.

Adjustments like those outlined in Table 10.6 may be small, but they're still important. Recall the energy levels in single-electron atoms (Eq. 10.2.25) were only dependent on n (the shell number), so there was a lot of **degeneracy**. However, many of these small adjustments are also dependent on ℓ (the orbital number) and j, so different orbital types and spin configurations can have *slightly* different energies. That means the degeneracy is broken in ℓ and j (see Figures 10.7 and 10.8). Breaking the degeneracy in orientation (m_j) requires an external magnetic field.

We also see the consequences in nature. For example, there is a famous spectral line of hydrogen called the "21 cm line" observed in interstellar clouds. Due to spin-spin coupling, the ground state of hydrogen (ψ^0_{10}) actually has two possible energy values that differ by

$$\Delta E = \Delta E_{\substack{\text{hf} \\ s=1}} - \Delta E_{\substack{\text{hf} \\ s=0}} = 5.874 \times 10^{-6} \text{ eV}.$$

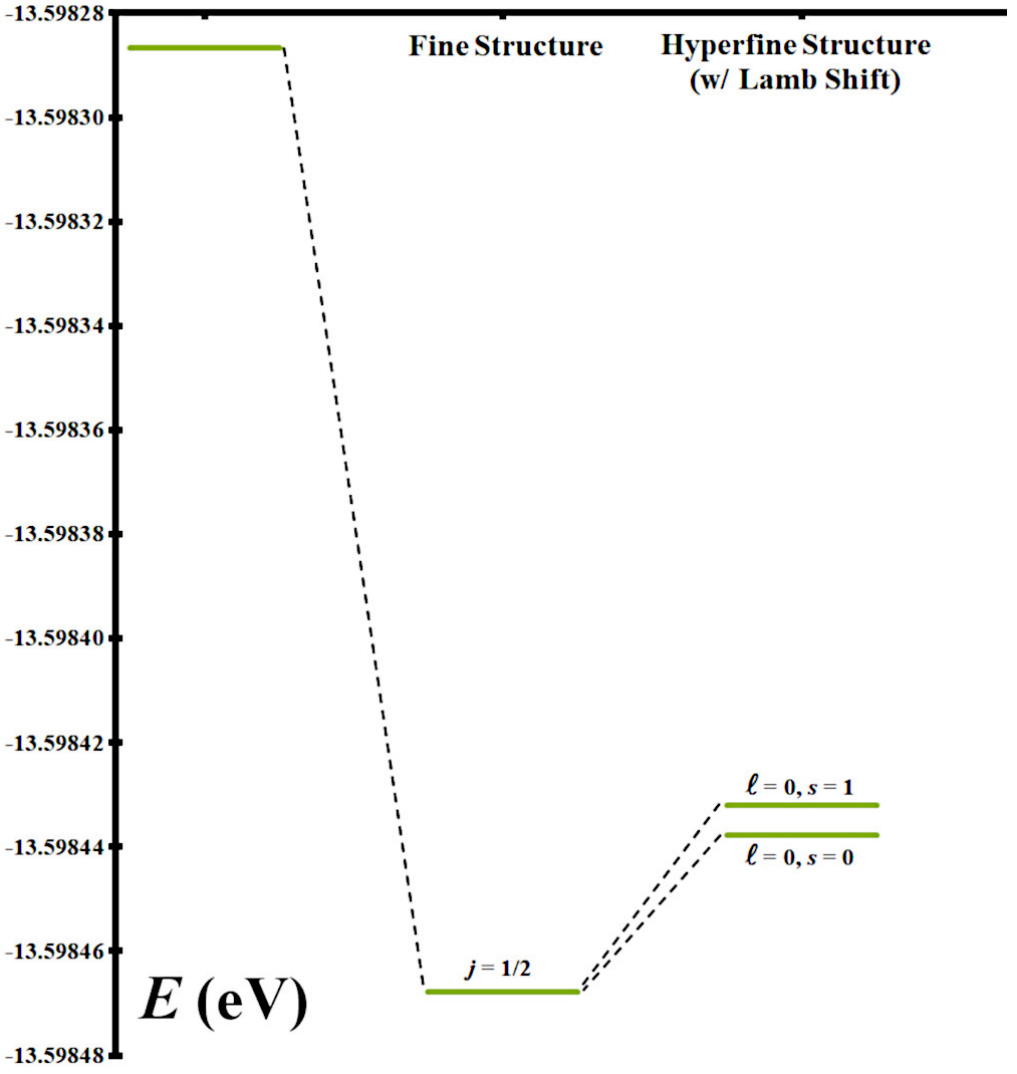


Figure 10.7: This is an energy level diagram for the first shell ($n = 1$) of the hydrogen atom. As you move to the right, sensitivity increases until all adjustments from Table 10.6 are included. A transition between the two hyperfine states results in the 21 cm spectral line observed in interstellar clouds.

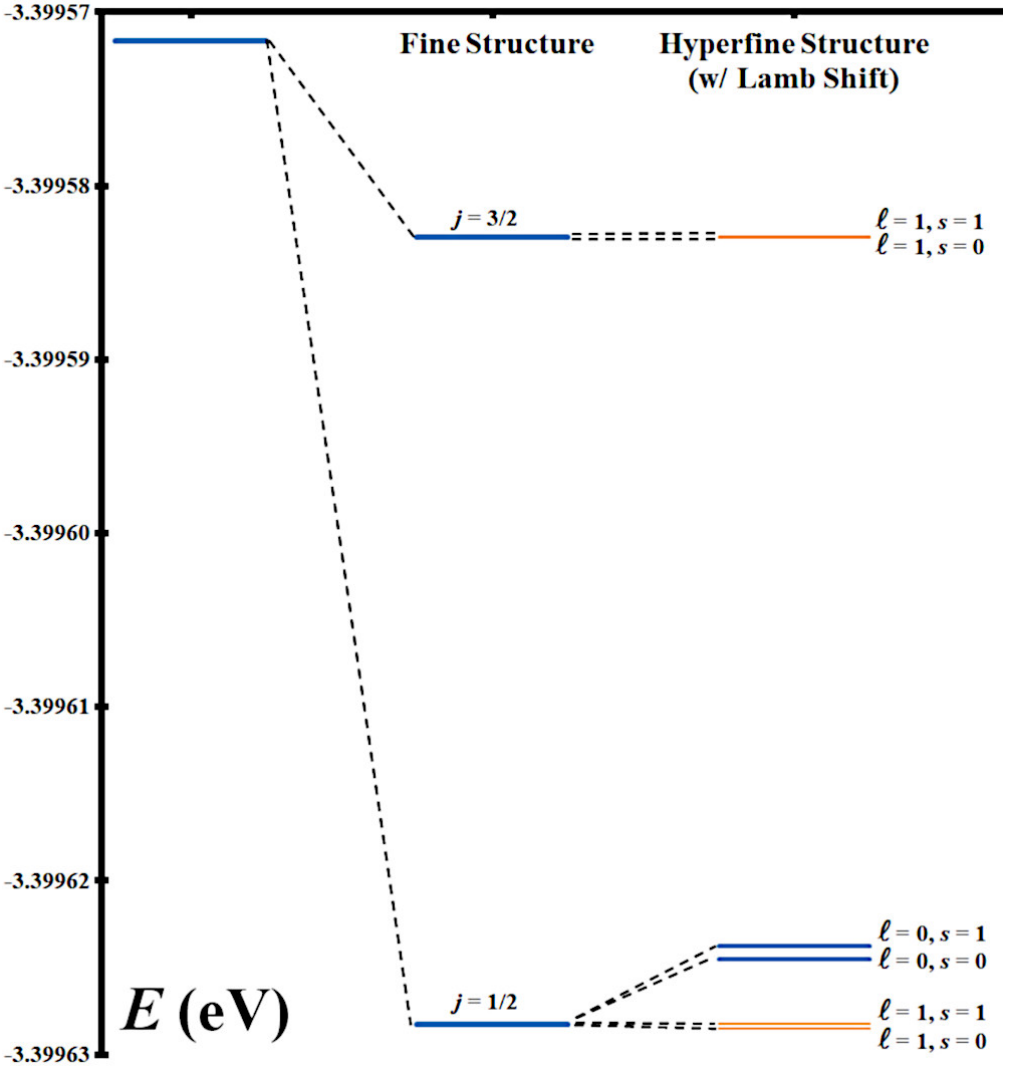


Figure 10.8: This is an energy level diagram for the second shell ($n = 2$) of the hydrogen atom. As you move to the right, sensitivity increases until all adjustments from Table 10.6 are included. Unlike in Figure 10.7, there are two fine structure states since j can be either $1/2$ or $3/2$. The p-orbitals ($\ell = 1$) have been shown in orange for clarity.

If the atom transitions from the higher to lower energy, then it will release a photon with a frequency of

$$f = \frac{\Delta E}{h} = 1420 \text{ MHz}$$

and a wavelength of

$$\lambda = \frac{c}{f} = 21.11 \text{ cm}$$

in the microwave range. Hyperfine transitions are also important in Cesium (atomic) clocks and refinement of nuclear material.

10.3 Multiple-Electron Atoms

The next logical question is "what happens when there is more than one electron?" Well, in short, things get complicated. Even neutral helium, which looks simple with only one extra electron, is difficult. It's called the **three-body problem** and any interpretation of the word "problem" is accurate in this instance. The three-body problem is infamous (even in classical mechanics) for often being analytically unsolvable. It tends to require numerical methods.

You should be cautious when carrying *anything* we learned about single-electron atoms into models of multiple-election atoms, but we'll see what we can do. First, the Hamiltonian for helium has five terms:

$$\mathcal{H} = KE + PE = KE_1 + KE_2 + PE_{\text{nuc},1} + PE_{\text{nuc},2} + PE_{1,2} \ ,$$

a kinetic energy for each electron and a potential energy for each interaction. It's the repulsion between the two electrons, $PE_{1,2}$, that causes all the trouble. Without it, the Hamiltonian is made of commuting parts (one for each electron), the solution would be separable ($\Psi = \Psi_1\Psi_2$) and we could carry *everything* over from single-electron atoms.

Unfortunately, the $PE_{1,2}$ term is just as significant as the others, so it cannot be ignored. Making the quantum substitutions for KE (Eq. 9.2.4) and PE (Eq. 10.2.1), we get

$$\mathcal{H} = -\frac{\hbar^2}{2m}\vec{\nabla}_1^2 - \frac{\hbar^2}{2m}\vec{\nabla}_2^2 - \frac{2q^2}{4\pi\epsilon_0 r_1} - \frac{2q^2}{4\pi\epsilon_0 r_2} + \frac{q^2}{4\pi\epsilon_0 \left|\vec{r}_2 - \vec{r}_1\right|}, \qquad (10.3.1)$$

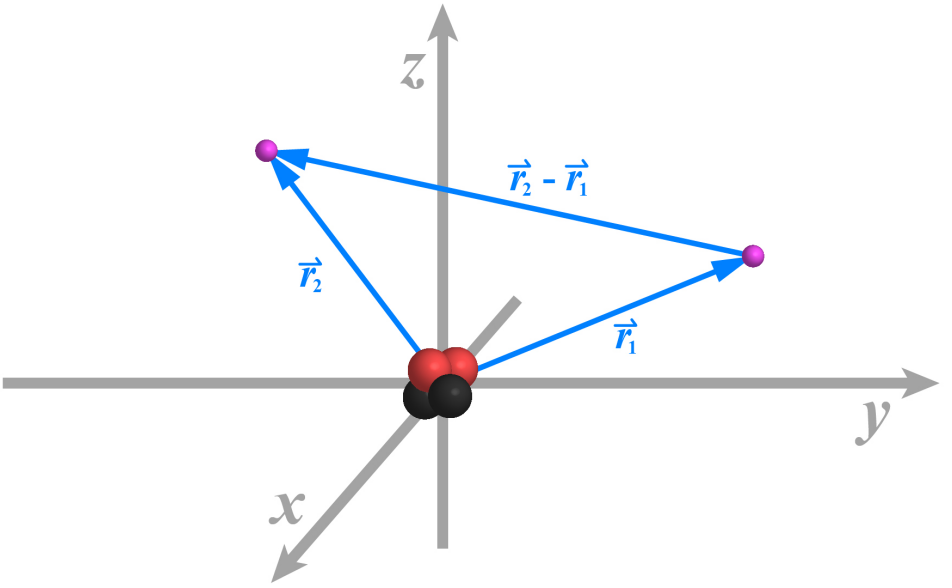


Figure 10.9: This is helium drawn as a three-body problem. The labels 1 and 2 correspond to each electron for use in Eq. 10.3.1.

where $m = m_e = 9.109 \times 10^{-31}$ kg and $q = q_p = +1.602 \times 10^{-19}$ C (see Figure 10.9). We saw in Eq. 10.2.32 the size of the nucleus (i.e. the value of Z) had no effect on the *shape* of the orbital. This is true even for multiple-electron atoms because the electron repulsion term,

$$PE_{1,2} = \frac{q^2}{4\pi\epsilon_0 \left|\vec{r}_2 - \vec{r}_1\right|}, \tag{10.3.2}$$

only depends on the distance between the electrons (i.e. it's independent of orientation). This means we can carry over the shapes in Figure 10.6. The orbitals are still s, p, d, f, etc. and are still determined by ℓ.

Electron repulsion terms (Eq. 10.3.2) can affect the *size* of an orbital and, therefore, it's energy. Speaking in general, for *any* atom larger than hydrogen ($Z \geq 2$), the Hamiltonian can be written as

$$\mathcal{H} = \sum_{l=1}^{Z} \left[-\frac{\hbar^2}{2m}\vec{\nabla}_l^2 - \frac{Zq^2}{4\pi\epsilon_0 r_l} \right] + \sum_{l=2}^{Z} \sum_{k=1}^{l-1} \left[\frac{q^2}{4\pi\epsilon_0 \left|\vec{r}_l - \vec{r}_k\right|} \right], \tag{10.3.3}$$

where the first summation represents all the kinetic energies plus interactions with the nucleus and the second summation represents all the electron

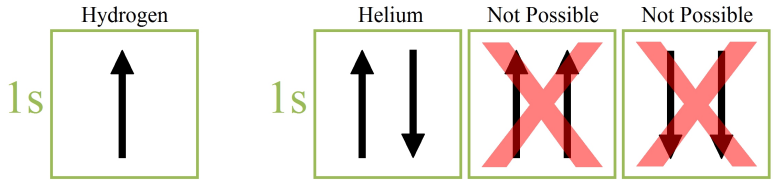


Figure 10.10: These are orbital diagrams for the ground state of hydrogen and helium. The arrows represent spin-up ($m_s = +1/2$) or spin-down ($m_s = -1/2$) electrons. The last two boxes show impossible cases for helium due to the Pauli exclusion principle.

repulsion energy. Now we'd like to know how these orbitals are filled as the atoms get larger. At any given time, the electrons could technically be in any of them. Statistically speaking though, they prefer to be in a state with as low an energy as possible.

Additionally, an electron cannot simultaneously occupy the same *total* quantum state as another electron. It's called the **Pauli exclusion principle**, named for Wolfgang Pauli, and applies to more than just the electron (see Appendix D for more details). Emphasis is put on the word "total" because electrons (all $s = 1/2$) can still have the same n, ℓ, and m_ℓ as long as m_s is different. Since $m_s = \pm 1/2$ for an electron, there can only be two electrons (one spin-up and one spin-down) in each orbital (i.e. each state given by n, ℓ, and m_ℓ). Figure 10.10 is an orbital diagram showing this phenomenon for hydrogen and helium.

Recall in Figure 10.8, there was a p-orbital lower than an s-orbital for $n = 2$. This is a phenomenon unique to hydrogen. Since the energy is affected by electron repulsion (Eq. 10.3.2), it breaks the degeneracy in ℓ without fine structure considerations. The base energy should now be written as $E_{n\ell}$ rather than just E_n. In all atoms larger than hydrogen ($Z \geq 2$), orbitals with the same n but a larger ℓ will have a higher energy (e.g. 2p is *always* higher than 2s, 3d is *always* higher than 3p, etc.). The same cannot be said when the values of n are different (e.g. 4p is *always* higher than 3d, yet whether 4s or 3d is higher depends on the atom). This occurs because the energy levels get closer together as they increase (see Figure 10.2). There is a set of rules for this called **Hund's rules**, but they have a ton of exceptions. I don't think any guideline with that many exceptions can really be called a "rule," so I'll show you a better way.

We also need to remember that it's not really the orbital that has energy. It's the electrons in those orbitals. Two different electrons in the same orbital

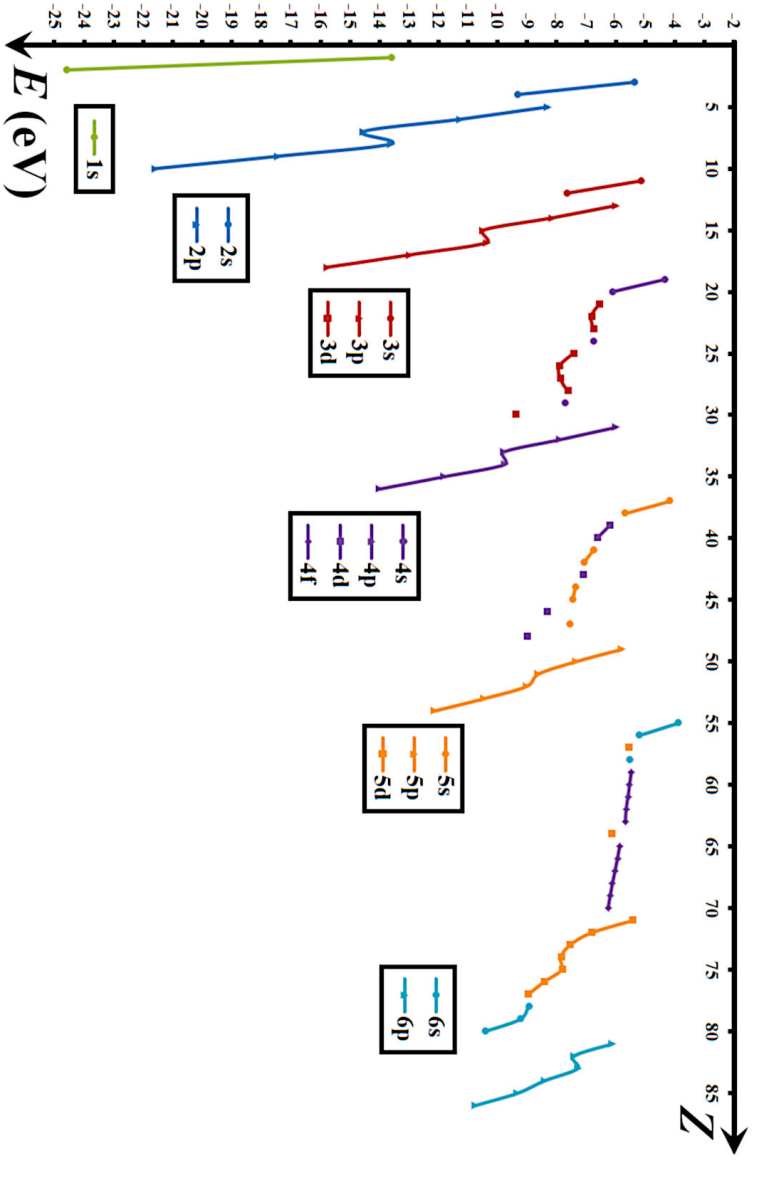


Figure 10.11: This graph shows the energies of the single outermost electron for each atom up to $Z = 86$ (the 6th row of the periodic table). They were found by determining the energy required to ionize each atom (i.e. remove that electron). Values of n (shell number) are indicated by color and values of ℓ (orbital type) are indicated by shape. You can clearly see where orbital types d and f make simple rules impossible.

can have two completely different energies. Furthermore, we usually only have experimental access to the outermost electrons (see Figure 10.11). The inner electrons are tightly bound, so transitions between them are *exceedingly* rare making experimentation difficult. In Figures 10.12 and 10.13, there is no scale on the energy axis because we're not exactly sure how much energy the 1s, 2s, 2p, 3s, or 3p electrons possess. A similar issue arises in Figures 10.14 and 10.15, so we've left the inner electrons out all together.

Periodic Table

All of this information about multiple-electron atoms and their orbitals gives us the ability to construct the **periodic table of elements**. As a bit of history, the periodic table was developed by Dmitri Mendeleev in 1870 CE, long before we even knew *for sure* that matter was made of atoms (although we suspected). It's not called the periodic table of "atoms" after all. You only have an "element" when there are enough of the same atom to make something exist on our scale of the universe. In other words, elements are macroscopic, but atoms are microscopic.

Mendeleev grouped elements into columns by similar chemical properties, then (assuming atoms existed) by atomic weight. At this point, the only thing we could know about atoms was that they were very small (as Democritus suggested in Section 9.1) because we couldn't see them under microscopes. Unfortunately, we didn't know exactly how small, let alone what they looked like. Atomic weight was measured relative to hydrogen, the lightest substance we had discovered, by balancing chemical equations.

Today, after almost two centuries of experiments, we know atoms are about $\frac{1}{10}$ nm in diameter (give or take). They are made of a nucleus (protons and neutrons) surrounded by a cloud of electrons. Most of the **atomic mass** (formally "atomic weight") is in the nucleus, but this is no longer a criterion for periodic table placement. Instead, we use the **atomic number**, Z, the number of protons. How the electrons are organized into orbitals (i.e. the **electron configuration**) determines the chemical properties of the element and, therefore, the columns (i.e. "groups") of the table. Unfortunately, the d-type and f-type orbitals often behave strangely, so this isn't as easy as it sounds.

The energy of electrons in d-type or f-type orbitals is significantly higher than the corresponding s-type or p-type, so the higher shells (determined

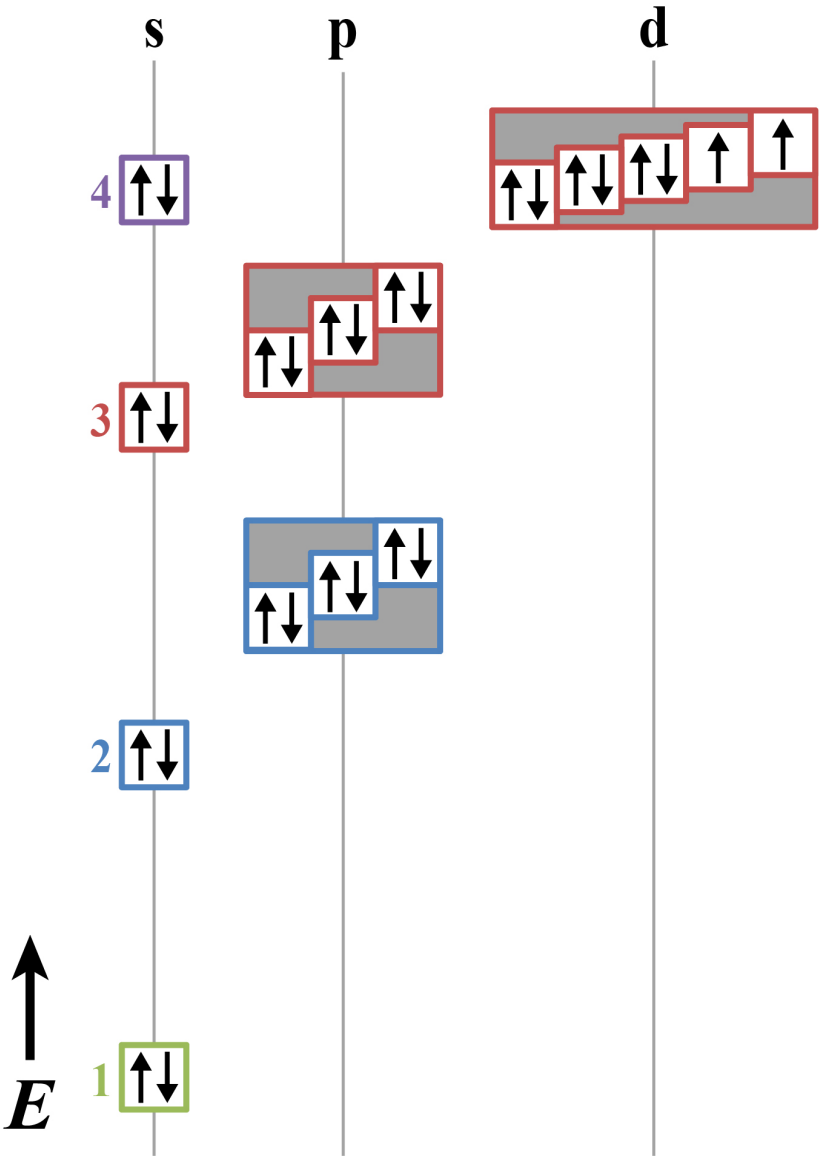


Figure 10.12: This is the orbital diagram for the ground state of Nickel ($Z = 28$). The arrows represent spin-up ($m_s = +1/2$) or spin-down ($m_s = -1/2$) electrons and the energy axis is not to any particular scale. Pairing opposite-spin electrons requires a bit more energy than lone electrons, so they tend to occupy every individual orbital (of each type) before pairing. Each box in each orbital-type is a single orbital and corresponds to a possible value of m_ℓ from Table 10.3.

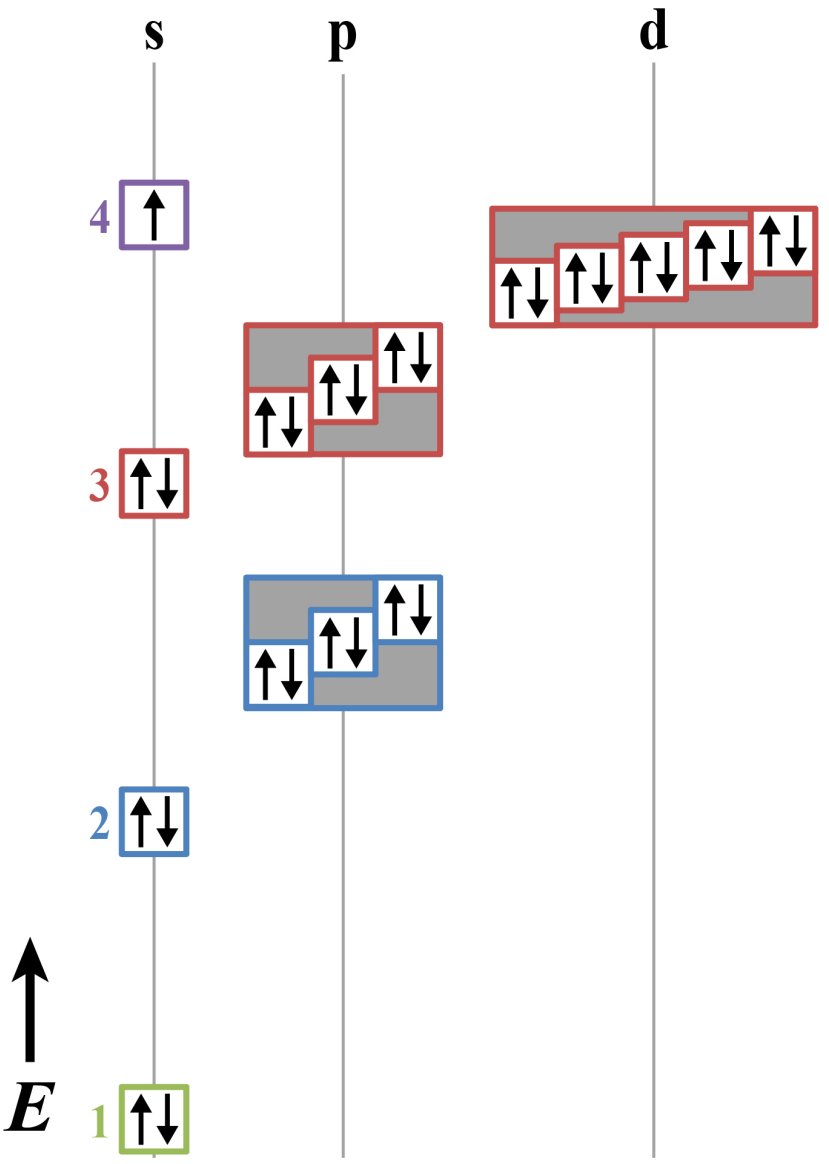


Figure 10.13: This is the orbital diagram for the ground state of Copper ($Z = 29$) similar to Figure 10.12. Since the nucleus is larger than Nickel's, it attracts the electrons more and all the orbitals are lower on the chart. However, 3d has more electrons in it, so it's attracted a little more bringing it lower than the 4s. As a result, a 4s electron falls into the remaining spot in 3d and the remaining 4s electron is very loose making copper a very good conductor of electricity.

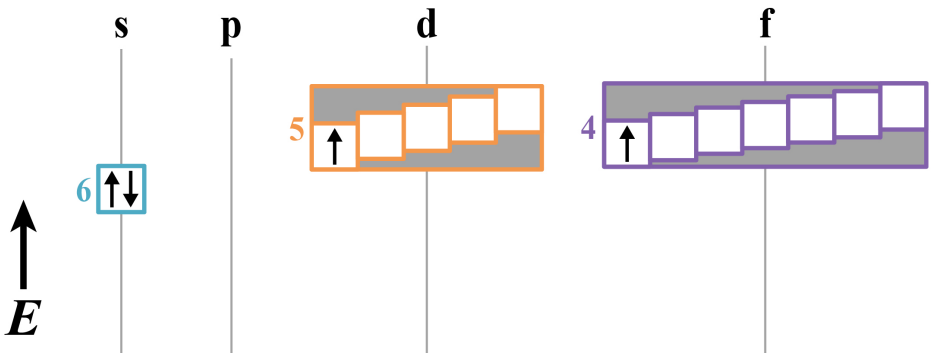


Figure 10.14: This is the orbital diagram for the ground state of Cerium ($Z = 58$) similar to Figures 10.12 and 10.13. We've included only the outermost electrons since we don't know much about those inner electrons anyway. The 4f and 5d electrons have almost exactly the same energy, so the 5d electron frequently oscillates between 5d and 4f.

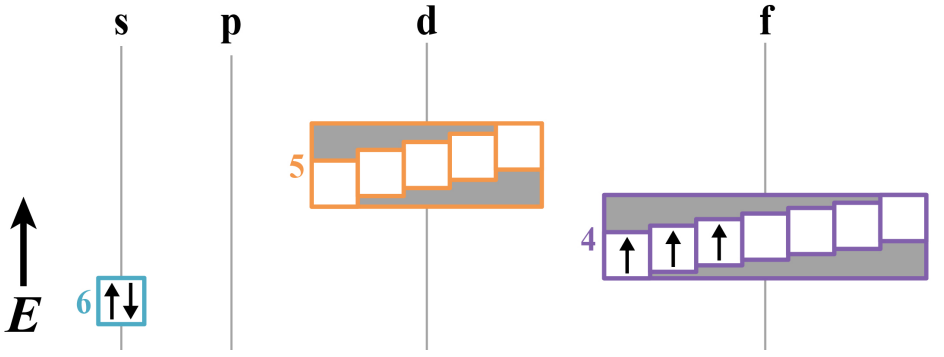


Figure 10.15: This is the orbital diagram for the ground state of Praseodymium ($Z = 59$) similar to Figure 10.14. Since the nucleus is larger than Cerium's, it attracts the electrons more and all the orbitals are lower on the chart. However, 4f has more electrons in it, so it's attracted a little more bringing it lower than the 5d. As a result, the 5d electron falls into a stable 4f state. Some electrons remain unpaired similar to Figure 10.12.

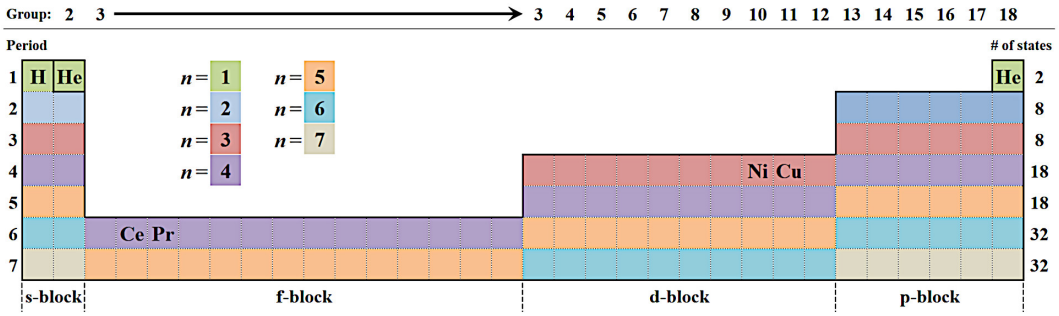


Figure 10.16: This shows how the periodic table is organized by n (shell number) and ℓ (orbital type). The elements shown in Figures 10.10 (hydrogen and helum), 10.12 (Nickel), 10.13 (Copper), 10.14 (Cerium), and 10.15 (Praseodymium) are also shown here for reference. Helium is sometimes shown two different places because it has the chemical properties of both groups.

by n) tend to bleed together. There are some examples of this in Figures 10.12, 10.13, 10.14, and 10.15. The orbitals fill in energy order from lowest to highest, not n or ℓ order. A guideline is given in Figure 10.16 in the shape of a periodic table. The figure only shows when each orbital type becomes important. For which orbital is on the outside, refer to Figure 10.11.

Rather than draw a full orbital diagram every time, we often simplify the electron configuration to a single line of text. Each term in the configuration is in the form

$$n\,(\text{Orbital Type})^{(\text{number of electrons})} \tag{10.3.4}$$

and you include one of these terms for each orbital being occupied. We know each orbital can only hold up to two electrons and we know how many orbitals each type has (Table 10.3), so

- s-types can hold $2 \times 1 = 2$,
- p-types can hold $2 \times 3 = 6$,
- d-types can hold $2 \times 5 = 10$, and
- f-types can hold $2 \times 7 = 14$.

This explains the number of boxes available for each orbital type in Figures 10.12, 10.13, 10.14, and 10.15. Some of these configurations can get a bit long, so we have a shorthand version. Usually, we're only interested in the

Table 10.7: These are the electron configurations of the few example atoms from this section. Noble gases (Argon and Xenon) have been used as shorthand.

Name	Symbol	Electron Configuration	Shorthand
Hydrogen	H	$1s$	$1s$
Helium	He	$1s^2$	$1s^2$
Nickel	Ni	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^2\ 3d^8$	[Ar] $4s^2\ 3d^8$
Copper	Cu	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^1\ 3d^{10}$	[Ar] $4s^1\ 3d^{10}$
Cerium	Ce	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^2\ 3d^{10}\ 4p^6\ 5s^2\ 4d^{10}\ 5p^6\ 6s^2\ 4f\ 5d$	[Xe] $6s^2\ 4f\ 5d$
Praseodymium	Pr	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^2\ 3d^{10}\ 4p^6\ 5s^2\ 4d^{10}\ 5p^6\ 6s^2\ 4f^3$	[Xe] $6s^2\ 4f^3$

orbitals on the *outside* of the atom. These correspond to the orbitals in that atom's row (or "period") of the periodic table (Figure 10.16), so we swap the other terms in the configuration with the noble gas symbol (far right of table) from the row above. A few examples are shown in Table 10.7.

This section might seem like it got a little "wordy" near the end since there wasn't much math we could do. For those of you who didn't bother to read any of it, here's a summary of the important bits:

- $n = 1, 2, 3, 4, \ldots$ is the shell number, which corresponds to a rough estimate of the energy of a collection of states. Each shell has n available orbital types (e.g. shell 2 has 2 orbital types: s and p).

- $\ell = 0, 1, 2, 3, \ldots$ is the orbital number related orbital angular momentum. The values correspond to orbital types s, p, d, f, etc. The shapes of s, p, and d are given in Figure 10.6.

- Electrons are spin-$\frac{1}{2}$ particles meaning the quantum number s is *always* $\frac{1}{2}$. It also means they only have two possible spin states: up ($m_s = +\frac{1}{2}$) and down ($m_s = -\frac{1}{2}$).

- Each orbital type (s,p,d,f) has a different number of orbitals (1,3,5,7).

- Each orbital can hold up to two electrons as long as their spin orientations, m_s, are opposite.

- Therefore, each orbital type (s,p,d,f) can hold a different number of electrons (2,6,10,14).

- However, since d-type and f-type orbitals are complicated, the number of spots in rows of the periodic table is not (2,8,18,32,50,72,98); but rather (2,8,8,18,18,32,32).

- With no external energy, electrons will fill orbitals from lower energy to higher energy with no exceptions. This order only *very loosely* corresponds to the order of n and ℓ, so thinking in terms of n and ℓ is *not recommended*.

- The periodic table is organized in order of atomic number (Z) from left to right, then grouped into columns according similar chemical properties.

- The size of the electron cloud shrinks as you move from left to right (in the periodic table) because the larger nucleus causes more attraction. The size grows as you move from top to bottom because more layers of electrons are added.

If you didn't read the paragraphs, I'd recommend you go back and do that in the future when you have time. Students often miss valuable information about the actual *physics* by only reading math, tables, and figures. Physics isn't in the math. It's in the language, concepts, and interpretation.

10.4 Art of Interpretation

In Section 9.2, we showed the only way to accurately represent subatomic particles was as waves of probability. If you *measure* an electron's position, then you will find it's located in only one place. Before the measurement though, you could only make predictions about the *chance* of finding it any particular place. That's the thing about statistics. It can be applied to just about anything, but the results aren't particularly profound.

What's the Problem?

Using statistics puts a limit on what we can discover about a physical system. For example, the statistical modeling of a gas as a collection of molecules gives us an idea of things like pressure, temperature, and entropy. In that case, the **microscopic** (i.e. small scale) only explains the **macroscopic** (i.e. large

scale). That's why we only tend to use statistics (in scientific theory) when everything else becomes impractical (e.g. when dealing with large numbers of objects).

However, as we saw in the examples in Sections 9.4 and 10.2, this is not the case in quantum mechanics where we apply statistics to *individual* particles. Why do we do that? We have no choice. As we saw in Section 9.1, when we try all the other mathematical tools, the whole model fails. Even when we add other behavior restrictions for no reason (e.g. Bohr's allowed orbits), the model falls short of explaining everything. The examples in Sections 9.4 and 10.2 had no real interpretation in them, so they were more like applied math than physics. The actual physics is a bit of an art and it can drive you a little crazy. Read forward at your own risk.

Ensemble of Particles

The interpretation that I think makes the most sense to people is that the wave function doesn't apply to a single particle, but an **ensemble of particles**. The idea is that, if you prepare say 10,000 identical experiments involving a certain kind of particle, then the wave function tells you how many of them will turn out a certain way. Recall the electron in the finite square well from Example 9.4.3:

- In Example 9.4.4, we found that there was a
  - 91.05% chance of finding that electron inside the well and a
  - 8.95% chance of finding that electron outside the well.
- According to this simple interpretation, if we prepared 10,000 identical wells just like this one and measured the position of the electron in each, then
  - 9,105 will show the electron inside the well and
  - 895 will show the electron outside the well.

The same happens for an electron in the p-orbital from Example 10.2.2. If you set up a bunch of these experiments and measure L_z, then 1/3 of them will come out $m_\ell = 0$ and 2/3 of them will come out $m_\ell = -\hbar$.

However, if you measure the position of the electron within an orbital, things get a little more visually interesting. The shapes of the orbitals are

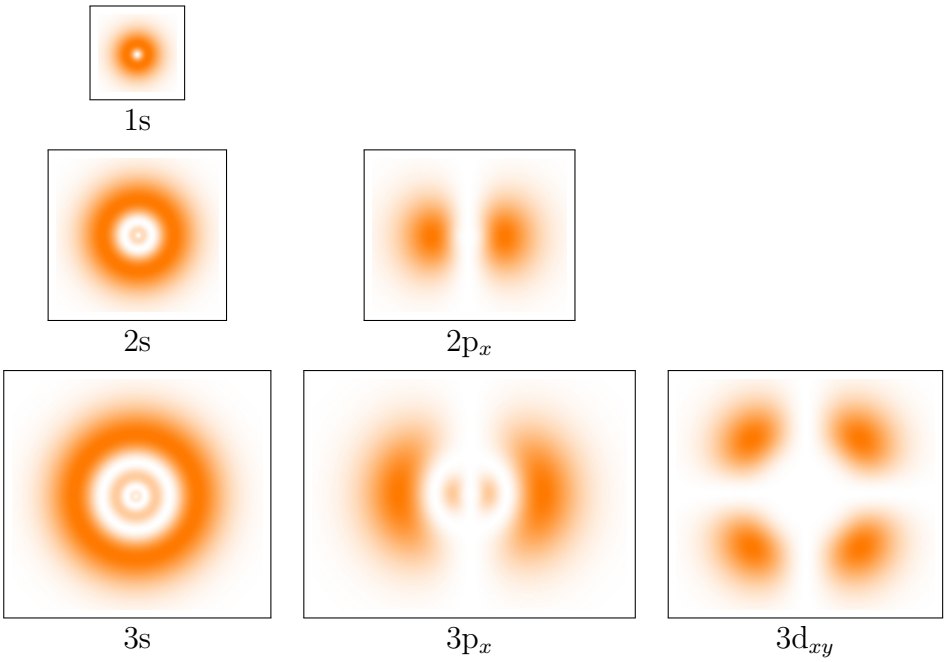


Figure 10.17: These are probability plots for a few orbitals in the hydrogen atom. Only the xy-plane cross section is shown for clarity. Orange pixels represent a measurement of the electron's position, so more concentrated orange means there is a higher probability.

given in Figure 10.6, but the electron is more likely to be found some places than others inside the shape. To account for that, we need to include $R_{n\ell}$ (Eq. 10.2.22) to get the full eigenstate (Eq. 10.2.27). Let's say you set up 10,000 electrons in identical hydrogen atoms, measure the positions of the electrons in each, and make a composite image of all 10,000. The result would be images like those in Figure 10.17.

All of this is certainly true about identical experiments, but does it *actually* mean the wave function doesn't apply to a single particle? This interpretation makes a lot of sense to people because it assumes it's just a problem of our ignorance. It says, somewhere underneath all this statistics, there is a **deterministic** theory (i.e. one where anything can be predicted as long as you know all the variables). Proponents argue there are just some **hidden variables** we can't yet measure. However, history has shown us, reality doesn't always lie in our comfort zone.

Bell's Inequality

In 1964, John Stewart Bell published a paper proving any *local* hidden variable theory was impossible. Let's say you have a neutral pion, π^0 (not to be confused with the negative pion, π^-, used in Example 7.4.1). The neutral pion is weird since it is its own antiparticle, so it decays into two photons,

$$\pi^0 \to 2\gamma, \tag{10.4.1}$$

about 98% of the time. This isn't very useful. Luckily, it decays into a photon and an electron-positron pair,

$$\pi^0 \to \gamma + \mathrm{e}^- + \mathrm{e}^+, \tag{10.4.2}$$

about 1.2% of the time. The electron (e^-) and positron (e^+) travel in opposite directions with opposite spins. Unfortunately, each has an equal probability of being the spin-up ($m_s = +1/2$) particle, so we would say the entangled pair is in the state

$$|0,0\rangle = \sqrt{\frac{1}{2}}\left|\tfrac{1}{2},+\tfrac{1}{2}\right\rangle^{\mathrm{e}^-}\left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle^{\mathrm{e}^+} - \sqrt{\frac{1}{2}}\left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle^{\mathrm{e}^-}\left|\tfrac{1}{2},+\tfrac{1}{2}\right\rangle^{\mathrm{e}^+} \tag{10.4.3}$$

since we don't know which is which. The Clebsch-Gordan coefficients are found from

$$|s,m_s\rangle = \sum_{m_{s1}+m_{s2}=m_s} C^{s_1,s_2,s}_{m_{s1},m_{s2},m_s}\,|s_1,m_{s1}\rangle\,|s_2,m_{s2}\rangle\,, \tag{10.4.4}$$

similar to those found using Eq. 10.2.51.

Now let's say each particle is headed toward its own spin detector, each of which capable of measuring along one of three unique orientations given by the unit vectors $\hat{a}$, $\hat{b}$, and $\hat{c}$. The orientation of each detector is chosen independently and at random for each successive measurement. If we *assume* the particles have definite spins the moment they are created (i.e. Eq. 10.4.3 just describes our lack of knowledge), then **Bell's inequality** states

$$\boxed{\left|\hat{a}\bullet\hat{c} - \hat{a}\bullet\hat{b}\right| \le 1 - \hat{b}\bullet\hat{c}\,.} \tag{10.4.5}$$

This must be true for *all* $\hat{a}$, $\hat{b}$, and $\hat{c}$ no matter how far apart the detectors; so one counterexample would show a contradiction. Setting

$$\left\{\hat{a} = \hat{x},\ \hat{b} = \hat{y},\ \text{and}\ \hat{c} = \frac{\hat{x}+\hat{y}}{\sqrt{2}}\right\}$$

gives us

$$\begin{aligned}\left|\hat{x}\bullet\left(\frac{\hat{x}+\hat{y}}{\sqrt{2}}\right)-\hat{x}\bullet\hat{y}\right| &\leq 1-\hat{y}\bullet\left(\frac{\hat{x}+\hat{y}}{\sqrt{2}}\right)\\ \left|\frac{1}{\sqrt{2}}-0\right| &\leq 1-\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} &\leq 1-\frac{1}{\sqrt{2}},\end{aligned}$$

which is *not* true. This leaves us with only two possibilities.

1. The universe is inherently non-local.

    - The measurement of the electron instantly determines *any* measurement of the positron. This is uncomfortable because all modern physics rest on the idea that information cannot travel faster than light.

2. There are no hidden variables.

    - Neither the electron nor the positron had a definite spin prior to the measurements. The particles were *physically* in a superposition of the two states (Eq. 10.4.3) until the measurement was made.

Bell's inequality has since been further generalized and many experiments have been done verifying all versions. As a result, the physics community has all but abandoned hidden variable theories.

Copenhagen Interpretation

Throughout the 1920s, Werner Heisenberg collaborated Niels Bohr in Copenhagen, Denmark. They were trying to come to some kind of agreement about what quantum mechanics was saying. In the end, they agreed on *almost* everything. Heisenberg gave a series of lectures in 1929 (and published a book in 1930) outlining the conclusions. He didn't coin the term **"Copenhagen interpretation"** until the 1950s while criticizing other interpretations. The term implies a level of historical formality that doesn't really exists. Still, I'll do my best at defining it.

We've already seen many of the principle ideas in the Copenhagen interpretation, but we'll include them again here in the interest of clarity. As of 1930, the description is as follows:

1. The wave function, $\psi(\vec{r}, t)$, *completely* describes the state of a system.
   (a) It is written as a superposition of *all* possible states weighted by the probabilities of each state.
   (b) It evolves *smoothly* in time according to Schrödinger's equation (Eq. 9.2.7) equation unless a measurement is made.
   (c) If a measurement is made, then the wave function *instantaneously* collapses to a stationary state of the observable being measured.
2. All quantum entities can display either particle properties, wave properties, or some combination of the two depending on the experiment being performed.
3. It is *not possible* to know *all* the properties of a system at the same time. Some observables will *always* be incompatible and the uncertainty principle (Eq. 9.3.31) must be applied in those cases.
4. The results of quantum physics must be consistent with classical physics in the macroscopic limit (i.e. large numbers of particles and/or large quantum numbers).

Since this list was made long before Bell's inequality (Eq. 10.4.5) was published in 1964, it doesn't do much "interpreting" really. Bohr felt quantum mechanics was useful in making predictions, but one should not read too far into it, which frustrated Heisenberg to no end.

However, since the publication of Bell's inequality (Eq. 10.4.5), the Copenhagen interpretation has developed into something much stronger and more suggestive. Some authors chose to call the **strong Copenhagen interpretation** by another name, but I don't see any reason to complicate matters any further. The strong additions are as follows:

1. The wave function, $\psi(\vec{r}, t)$, represents the *physical existence* of the system.
2. If a particle is in a stationary state of an observable, then it will have a definite value for that observable.

- If that observable is measured, the particle will display that value.

3. If a particle is not in a stationary state of an observable, it will *exist* as a superposition of those stationary states.

   - If that observable is measured, the particle will *instantaneously* and *randomly* collapse into a single stationary state and display the value of that state.
   - The randomness of that collapse is weighted by the probabilities contained in the wave function.
   - **It isn't just that we can't predict them. It's that the particle doesn't have them.**

The difference between what we can predict and what we actually measure is tricky business, but both have equal footing in physical reality.

Particles vs. Waves

The best way to make sense of all this craziness is with context. A very famous thought experiment by Richard Feynman might help with this. It's a generalization of the **double-slit experiment** Thomas Young used in 1801 to show that light was a wave. The purpose of the thought experiment is to distinguish between predictions and measurements when it comes to quantum particles like electrons. It will also more clearly define what we mean by particle properties and wave properties.

We're going to set up three similar experiments following to the setup shown in Figure 10.18. The experiments will proceed as follows:

1. Subject: *Bullets*

   Source: Machine gun

   Slit plate: Metal armor

   Detector: Box of sand

   Assumptions: Bullets (and armor) are indestructible.

   - As the bullets pass through the slit plate, they ricochet off the armored walls in all directions. The ones that make it through, will make their way toward the box of sand and stop. After an

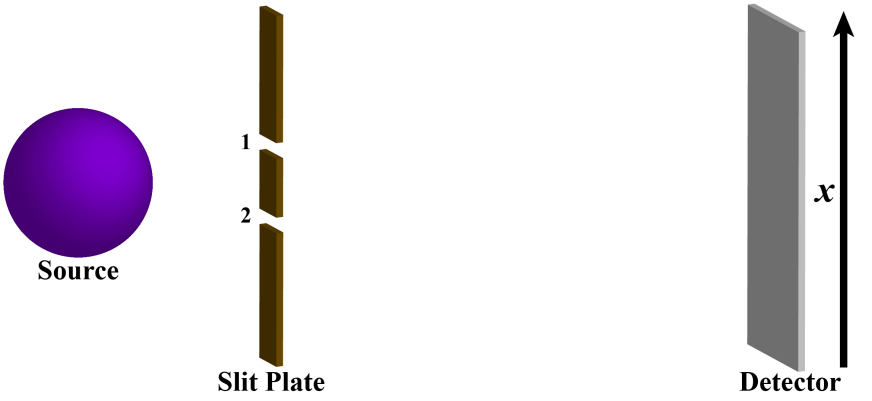


Figure 10.18: This is the basic experimental layout for Feynman's double-slit thought experiment. Position, x, along the detector is measured from the bottom edge. The openings in the slit plate are labeled 1 and 2 for reference.

hour, we stop the experiment, count the bullets at each point in the box, and repeat the experiment several times to get an average. We're measuring bullets per hour, which is a kind of **intensity**.

2. Subject: *Water*

   Source: Piston

   Slit plate: Wood

   Detector: Chain of floating buoys

   Assumptions: Piston and buoys can only move vertically.

   - As the piston moves, surface waves are created on the water that move in all directions. The ones that make it through the slit plate, will make their way toward the buoys and cause them to bounce. We measure the maximum displacement of each buoy for an hour (i.e. the **amplitude**) and take an average.

3. Subject: *Electrons*

   Source: Filament

   Slit plate: Tungsten radiation shielding

   Detector: Chain of Geiger counters

Assumptions: Geiger counters don't miss electrons.

- While the filament is on, electrons are released in all directions. The ones that make it through the slit plate, will make their way toward the Geiger counters and cause them to click. We count the clicks from each Geiger counter for an hour, stop the experiment to record, and repeat the experiment several times to get an average. We're measuring electrons per hour.

We'll run through each experiment three different ways: once with both slits open, once with only slit 1 open, and once with only slit 2 open. This will allow us to examine their true behavior.

The ultimate result of each experiment is going to be a comparison between how the detector pattern looks from two open slits (labeled I_{12}) and how we *expect* it to look based on the two single-slit patterns (labeled I_1 and I_2). If the subject of the experiment is a particle, then they will just build up independently and the two single-slit patterns will simply add. In terms of **intensity** at each value of x on the screen in Figure 10.18, that can be written as

$$I_{12}(x) = I_1(x) + I_2(x)\,. \tag{10.4.6}$$

If the subject of the experiment is a wave, then it's the disturbances (i.e. amplitudes) of the wave that add. In terms of **amplitude** at each value of x, that can be written as

$$A_{12}(x) = A_1(x) + A_2(x)\,. \tag{10.4.7}$$

Since intensity is proportional to the square of the amplitude,

$$\begin{aligned} I_{12} &\propto (A_{12})^2 \\ I_{12} &\propto (A_1 + A_2)^2 \\ I_{12} &\propto (A_1)^2 + (A_2)^2 + 2A_1A_2\cos(\varphi_0)\,, \end{aligned}$$

having used the law of cosines in the last step. We also know $I_1 \propto (A_1)^2$, $I_2 \propto (A_2)^2$, and φ_0 is the phase difference between the two waves; so

$$I_{12}(x) = I_1(x) + I_2(x) + 2\sqrt{I_1(x)\ I_1(x)}\cos\left(\frac{2\pi d}{\lambda}\frac{x}{z}\right). \tag{10.4.8}$$

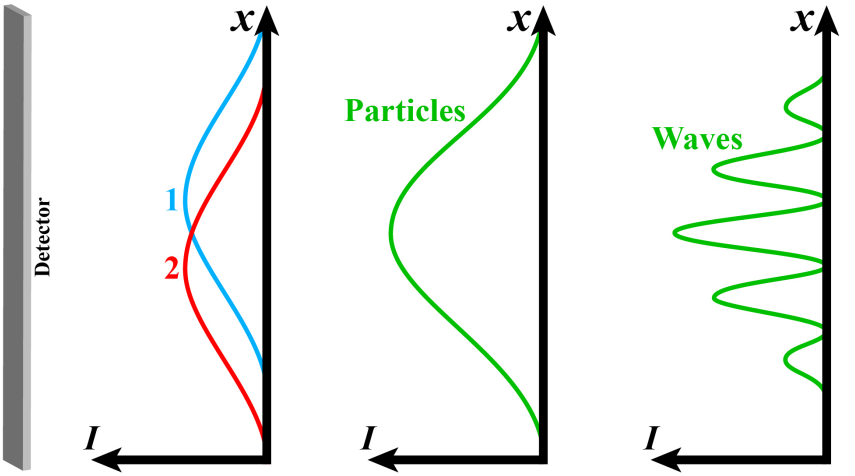


Figure 10.19: The first graph shows the intensity from each individual slit when the other is closed. The second graph shows the intensity when both slits are open if you're firing *particles* (e.g. bullets). The third graph shows the intensity when both slits are open if you're firing *waves* (e.g. water).

where λ is the wavelength of the wave, d is the distance between the slits (comparable to λ), and z is the distance between the slit plate and detector ($z \gg x$).

The graphs for I_1 and I_2 will look very similar for both particles and waves (due to the behavior of waves passing through a single small opening). However, as you can see in Figure 10.19, the graphs for I_{12} look very different.

1. Bullets from one slit don't "interfere" with bullets from the other slit, so they behave like particles showing the pattern in the second graph in Figure 10.19.

2. Water waves are a different story. As the waves exit the two slits, they spread out and overlap. The water must respond to both *simultaneously*, which is what we call **interference**. By the time the waves get to the chain of buoys, some parts are adding together and some are canceling out. This results in the third graph in Figure 10.19.

Both experiments have shown exactly what we would expect and we now have a basis from which to judge electrons.

According to classical physics, electrons are particles, so there are no partial electrons and they must travel a certain path (i.e. they take *either*

slit 1 *or* slit 2, but never both). Based on this, we expect the electron's detector pattern to match the one for bullets (see Figure 10.19). We even counted electrons just like we counted bullets: hits at each position x per hour (i.e. $I_{12} = N_{12}$).

3. Electrons are tricky beasts though. When we perform the experiment, our measurements match the pattern for waves (i.e. the third graph in Figure 10.19). The experiment says electrons are waves.

By this point in the book, you're already well aware of this. According to Section 9.2, they're probability waves, but what does that *actually* mean?

If we're counting electrons like we count bullets, then let's take another look at bullets. We'll use the total number of bullets fired per hour to normalize the intensity curve:

$$I_{12}(x) = N_{12}(x) \quad \Rightarrow P_{12}(x) = \frac{I_{12}(x)}{N_{\text{total}}} = \frac{N_{12}(x)}{N_{\text{total}}},$$

where $P_{12}(x)$ is the probability of getting a bullet at x when both slits are open. We're really just measuring probability. If electrons are interfering like waves, then we'll need an analog to amplitude such that its square is the probability. We'll call it a **probability amplitude** and the electron's wave function,

$$\psi(x) = \langle x | \, | \psi \rangle \, ,$$

conveniently fits the criteria. This allows us to use the same kind of math for electrons. Unfortunately, particle wave functions are complex (i.e. containing both real and imaginary parts) and the probabilities (Eq. 9.3.8) are complex squares,

$$P(x) = \| \langle x | \, | \psi \rangle \|^2 \, ,$$

so we can't make any physical sense of it like we could for water waves (i.e. all analogies stop here). In the case of the double-slit experiment, the probability of an electron arriving at x is

$$\begin{aligned} P_{12}(x) &= \left\| \sum_{\text{slits}} \langle x | \, | \text{slit} \rangle \, \langle \text{slit} | \, | \psi \rangle \right\|^2 \\ P_{12}(x) &= \| \langle x | \, | 1 \rangle \, \langle 1 | \, | \psi \rangle + \langle x | \, | 2 \rangle \, \langle 2 | \, | \psi \rangle \|^2 \\ P_{12}(x) &= \| \psi_1(x) + \psi_2(x) \|^2 \end{aligned}$$

where the probability amplitudes ψ_1 and ψ_2 act similar to the amplitudes A_1 and A_2 in Eq. 10.4.7.

The next logical question: "What is actually interfering?" A good guess would be that electrons passing through slit 1 are interfering with those passing through slit 2 (i.e. how the water behaves). We can easily test this by cooling the filament source until one electron is released at a time. Without another electron, there certainly can't be interference, right? Wrong! If you perform the experiment this way, then it takes longer, but you'll *still* get the third graph from Figure 10.19. There is only one possible conclusion is the electron interferes with itself or, more bluntly, a single electron can pass through both slits. It must pass through both simultaneously, otherwise there would be no interference pattern.

If an electron can pass through two slits at the same time, then we should be able to check for that! We'll set up a light source and a couple sensors next to the slit (see Figure 10.20). If the light is scattered, then one of the sensors will activate and we'll know an electron went through that particular slit. Performing this version of the experiment results in a surprise: each observed electron passes through only one slit. However, now that we've observed which slit each one passes through,

$$\begin{aligned} P_{12}(x) &= \|\langle x|\,|1\rangle\,\langle 1|\,|\psi\rangle\|^2 + \|\langle x|\,|2\rangle\,\langle 2|\,|\psi\rangle\|^2 \\ P_{12}(x) &= \|\psi_1(x)\|^2 + \|\psi_2(x)\|^2 \\ P_{12}(x) &= P_1(x) + P_2(x) \end{aligned}$$

and the detector pattern matches the one for particles (see Figure 10.19).

> **When we look for them to be particles, they behave like particles. When we don't, they behave like waves.**

Prior to its detection at the slit plate, the electron was in superposition of slit 1 *and* slit 2. The act of observing the electron's path forced the electron to collapse into a state of slit 1 *or* slit 2, but not both. It would seem particles don't like to be watched by experimenters.

As with every other thought experiment in this book, this one has limits. For double-slit diffraction to be noticeable, the slit size and separation both have to be comparable to the wavelength of the wave. In the case of visible light, the wavelength is $\approx 10^{-7}$ meters ($\approx 0.1\ \mu$m), so slit scales can't be much larger than 10^{-5} meters ($10\ \mu$m). Electron wavelengths tend to be $\approx 10^{-10}$

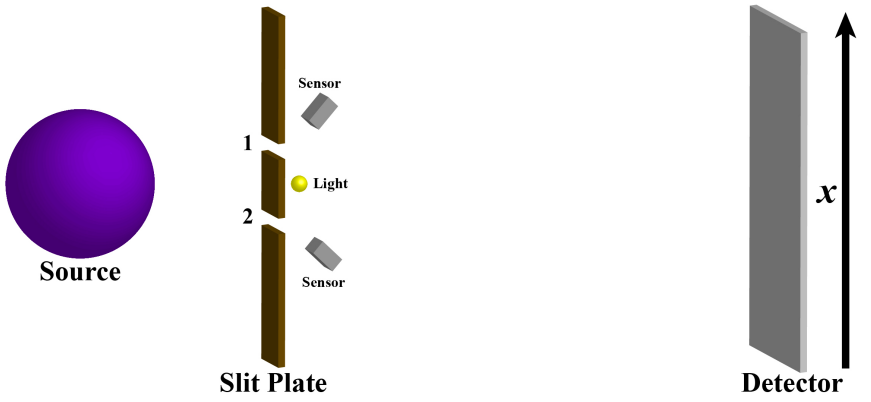


Figure 10.20: This is an experimental layout for Feynman's double-slit thought experiment (like Figure 10.20), but with a light source and some sensors added to detect which slit is being used by which electrons. Position, x, along the detector is measured from the bottom edge. The openings in the slit plate are labeled 1 and 2 for reference.

meters (≈ 0.1 nm), which is 1000 times smaller than visible light. This means the slit scales also need to be about 1000 times smaller or $\approx 10^{-8}$ meters (≈ 10 nm). That was impossible for decades after Feynman's proposal, but in 2012 the experiment was finally done in real life and the results given here have been confirmed. We can no longer treat this as *just* a thought experiment.

Macroscopic vs. Microscopic

At the beginning of this section, we mentioned the terms **macroscopic** meaning "large scale" and **microscopic** meaning "small scale." We contrast the two often in science (e.g. when discussing elements vs. atoms in the periodic table) and quantum mechanics is no exception. In fact, quantum mechanical weirdness requires we be *extra* careful with what we mean by the terms. With only a quick glance, it would seem microscopic particles are somehow *aware* of the macroscopic world and change their behavior accordingly, which is absurd. We need to delve into this a little deeper.

Let's take another look at the Feynman double-slit experiment (see Figure 10.18). When we detected which slit each electron passed through by shining light on it. Well, light also displays wave-particle duality since we can say it's made of individual photons. Those photons are what scatter off electrons to indicate their location.

We made the "measurement" using a microscopic tool: the photon.

In order to detect *all* the electrons passing through the slit plate, we need there to be *a lot* of photons. If we turn the brightness down, then we'll have fewer photons and it's possible some of the electrons make it through undetected. The detected electrons will behave like particles, the undetected ones behave like waves, and you get a detector pattern somewhere between particles and waves (see Figure 10.19).

The electron is only aware of the interacting photon, not the experimenter.

There is some mechanism in the interaction between the photon and electron that changes which properties the electron displays (and the photon, for that matter). Unfortunately, we have no idea of the nature of that mechanism.

The point is photons hit electrons all the time without the need of an experimenter and the same thing happens: the electron displays a single position. Sorry to burst any of your bubbles, but:

A "measurement" doesn't require a conscious mind.

It just seems to require a certain kind of interaction. I say "a certain kind" because not all interactions collapse the wave function, only some do, and we don't have a clear definition of either category. We probably should have used a different word when quantum mechanics was in its infancy, but now we're stuck with it.

You might be wondering though: "What's the deal with wave function collapse?" It's a good question to ask and we'll make sense of it by returning to a simple model: the infinite square well (Example 9.4.1). If an electron is in the stationary state,

$$\psi_3(x,t) = \sqrt{\frac{2}{a}} \sin\left(\frac{3\pi}{a}x\right) e^{-i\left(5.142\times 10^{15}\,\frac{\text{nm}^2}{\text{s}}\right)t/a^2},$$

found using Eq. 9.4.8, then it will have a definite energy,

$$E_3 = 3.385\,\frac{\text{eV nm}^2}{a^2},$$

found using Eq. 9.4.5 where a is the width of the well. However, it will not have a definite position because the observable x is incompatible with the

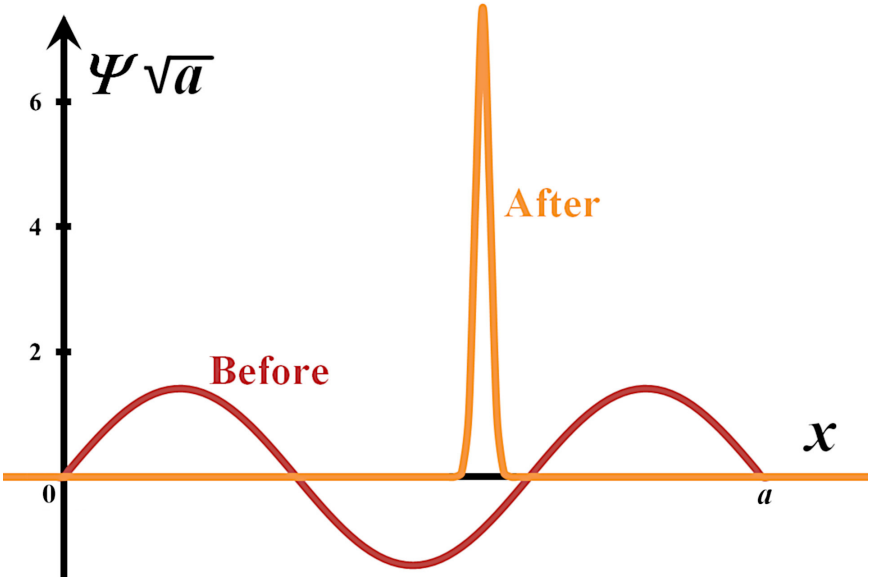


Figure 10.21: This graph represents the state and electron in an infinite square well *before* and *after* a measurement is made of it's position, x. Prior to the measurement, it's in a stationary state of the Hamiltonian, $\mathcal{H}$. After, it's in a stationary state of the position, x.

Hamiltonian (i.e. $[\mathcal{H}, x] \neq 0$). Before x is measured, the electron is in a superposition of all possible values of x inside the well. This is easily seen in Figure 10.21 because the stationary state, ψ_3, is written in the position basis. When we measure x, the electron must only be in one place since it's a point charge. It must collapse into a stationary state of x, rather than a stationary state of $\mathcal{H}$, as shown in Figure 10.21.

Now that the position is definite, the energy is not. The electron exists as a superposition of all the energy states (see Eq. 9.4.9) until we try to measure its energy again, at which point it will change to one of those. This is why we have the uncertainty principle (Eq. 9.3.31) and it occurs any time there's an interaction that determines the value of an observable, experimenter or not.

> **The more definite some observables become, the less definite some others become.**

Furthermore, you can never know any property *exactly*. This is why, even after a measurement of x, Figure 10.21 shows the electron is around $0.6a$

give or take a little (indicated by the width of the "spike"). It's not just an experimental problem. It's a physical one.

Bridging the Gap

Part of the Copenhagen interpretation requires the results of quantum physics be consistent with classical physics in the macroscopic limit, so we can't keep them separate forever. After all, this is really just one universe. To bridge the gap, we're going to have to ask ourselves a tough question: "Where *is* the macroscopic limit?" Where does the microscopic world end and the macroscopic world begin? This is not an easy question to answer and, as far as I know, no one has a good one.

We'll start our attempt at an answer with a famous thought experiment called **Schrödinger's cat**. The idea is that a cat is placed in a sealed box with no windows. Also in the box is a poison activated only by the random decay of a radioactive material.

- If an atom in the material decays, then the cat dies.
- If an atom in the material doesn't decay, then the cat lives.

Immediately after the box is sealed, the atom is in a superposition of decayed and not decayed. Since the cat is linked to the atom, it is in a superposition of dead and alive. You'd write it something like:

$$\psi = \sqrt{\frac{1}{2}}\,\psi_{\text{dead}} + \sqrt{\frac{1}{2}}\,\psi_{\text{alive}},$$

where the coefficients imply equal probability. This experiment suggests the cat isn't in a definite state, which seems preposterous.

We've dealt with paradoxes like this before in Section 7.7. Paradoxes are not something that really exist in nature, but they do exist *on paper* for one of two reasons:

1. A false assumption given the nature of the model being used, or
2. That we've stepped beyond the scope of the model.

In the case of Schrödinger's cat, it's the first reason. In particular, what activates the poison? It doesn't just happen magically. If it's not the experimenter and it's not the cat, then there must be something in the box

that detects the decay. A Geiger counter would suffice! However, isn't that a measurement? There are two possibilities:

- The Geiger counter detects the radiation, activates the poison, and the cat dies. The superposition state of the atom collapses into a single state.
- The Geiger counter doesn't detect the radiation, the poison is inactive, and the cat lives. The atom remains in superposition state.

There is no superposition for the Geiger counter and, therefore, no superposition for the cat.

This brings up a good point though. A macroscopic object *never* exists in a superposition, so what happened? Certainly the cat is made of quantum particles, so why doesn't it behave that way?

> **Macroscopic things are always either particle-like or wave-like, but never both.**

We don't have this quite figured out yet, but allow me to *speculate* for a moment. We know the cat is made of quantum particles, but how many? Just counting atoms, that would be about 10^{27} (order of magnitude approximation). Those atoms are mostly hydrogen, oxygen, and carbon, so the number of subatomic particles could easily be about 10^{28}. That's a lot of particles!

Those particles interact quite a lot and I'd imagine a fair portion of those interactions could be considered "measurements," so those wave functions must be collapsing a lot.

> **Wave-particle duality gets lost in the large particle system.**

Recall what happened in Section 9.2 when we tried to argue a single electron was just charge smeared out across an orbit? It failed. However, the billions of electrons on a charged surface certainly behave that way. If you modeled the billions of individual electrons as probability waves and used a big computer to simulate the whole process, then you just wont see any of the wave properties. Some physicists call this **quantum decoherence**.

This explanation looks great until you remember that not all macroscopic things are particle-like. Huge collections of electrons might lose their wave properties, but huge collections of *photons* lose their particle properties.

Light behaves like a *wave* on the large scale. This could have something to do with mass (i.e. electrons have mass and photons don't), but no one knows for sure. I just don't think you can ask where the macroscopic world begins because it's more a continuous gradual process. The more particles there are and the more space they take up, the less and less duality there appears to be. It's always there, but one or the other just becomes significantly more dominant.

Interpretations or not, quantum mechanics is weird and crazy. It can make even the most skilled physicist pull out there hair just thinking about it. We use it though because it works. It can make incredible predictions that would have been impossible to make without it and we've performed countless *real* experiments verifying its principles. In the future, it may turn out that quantum simultaneously applies to every copy of a particle in an infinite multiverse (i.e. we don't know which particle we have in our universe until we "look"). It might even turn out that our universe is inherently non-local allowing for other hidden variable theories. Unfortunately, most of us will just have to wait and see.

Appendix A

Numerical Methods

A.1 Runge-Kutta Method

The fourth-order Runge-Kutta method (or sometimes just the Runge-Kutta method) was developed by German mathematicians Carl Runge and Martin Wilhelm Kutta around 1900. It is a method of integrating first-order differential equations numerically. It is particularly useful in the cases that are not solvable analytically, which arises quite often in Lagrangian mechanics (discussed in Chapter 4) as well as other fields.

We begin with the initial condition $y(t_0) = y_0$ and then move forward step by step using

$$\left\{\begin{array}{lcl} y_{n+1} & = & y_n + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right)\Delta t \\ t_{n+1} & = & t_n + \Delta t \end{array}\right\} \tag{A.1.1}$$

where

$$\left\{\begin{array}{lcl} k_1 & = & \dot{y}(t_n, y_n) \\ k_2 & = & \dot{y}\left(t_n + \frac{1}{2}\Delta t, y_n + \frac{1}{2}k_1\Delta t\right) \\ k_3 & = & \dot{y}\left(t_n + \frac{1}{2}\Delta t, y_n + \frac{1}{2}k_2\Delta t\right) \\ k_4 & = & \dot{y}(t_n + \Delta t, y_n + k_3\Delta t) \end{array}\right\} \tag{A.1.2}$$

and Δt is constant and called the iteration step.

By now, you may have noticed this method applies only to first-order differential equations and those that occur in Chapter 4 are second-order.

This is not a problem because, with a little algebraic manipulation, higher-order equations can be written as a system of first-order equations.

Example A.1.1

Turn the following second-order differential equation into a set of first-order equations.

$$\ddot{y} + 2\dot{y} + 4y = 25$$

Note: This equation has no physical context what-so-ever.

- First, we'll solve this equation for $\ddot{y}$ and we get

$$\ddot{y} = 25 - 2\dot{y} - 4y.$$

- Second, we'll define a new quantity v as the first derivative of y. This results in a set of

$$\left\{\begin{array}{rcl} \dot{y} &=& v \\ \dot{v} &=& 25 - 2v - 4y \end{array}\right\},$$

  which is a system of two first-order differential equations.

- Yes, it creates an extra variable, but it's necessary if you intend to solve the original equation using the Runge-Kutta method.

- This method can be applied to even higher order equations, but for third-order there will be three equations and for fourth-order there will be four and so on. Furthermore, y can be either a scalar or a vector quantity.

Contrary to its general appearance, the Runge-Kutta method of integration is not impervious. The accuracy of the method depends on two things:

- The initial values

- The iteration step

Taking another look at Eqs. A.1.1 and A.1.2, we can see that the value of $y(t)$ will not change with each iteration if $y(t_0)$ and $\dot{y}(t_0)$ are both zero. The method will always result in a zero. However, taking an extra derivative of your function to include an extra initial value will usually solve this problem. Just remember that doing so will add an extra set of integration.

This brings us to the iteration step. As long as the iteration step is sufficiently small, the graphical result will be accurate. How small is "sufficiently small?" Well, that will depend on your differential equation(s) and the level of desired accuracy. There are also times when you may want to relax the condition that the iteration step be constant. This is called adaptive iteration and involves knowing a little something about your function. Using terrain as an analogy, you should make your step smaller when passing through erratic mountainous regions to guarantee accuracy in those areas and you can make it larger when passing through smooth countryside to increase speed.

A.2 Newton's Method

Suppose you have a transcendental function (i.e. it "transcends" algebra) for which you need to find an inverse or simply want to find the solution. I realize you could probably throw this into a graphing calculator or some computer program (e.g. Mathematica, MAT LAB, etc.), but haven't you even wondered what those tools are doing to find those solutions? It's important to understand how these tools work on some level because you'll want to make sure they're doing it correctly for your application. A tool is only as good as its user.

Newton's Method is a good approach for a situation such as this one. First, you set your equation equal to zero,

$$f(x) = 0, \tag{A.2.1}$$

that way all relevant information about the equation is together (no matter how nasty it looks). Now your solutions, x, are zeros of f. Second, you'll find its first derivative, $f' = df/dx$. Newton's method also requires you start with a guess, x_0, but don't worry too much about it.

- The closer your guess is to the solution, the less time this method takes.
- However, as long as you're closer to the desired solution than any other solution, the method will *always* work.

Once you have a guess, you step progressively closer to the solution using

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \tag{A.2.2}$$

where n is a whole number (i.e. $n = 0, 1, 2, 3, \ldots$) and $f' = df/dx$ is the first derivative with respect to x.

Example A.2.1

Solve $3e^x + xe^x = 9$ for x.

- First, we set the equation equal to zero to find f:

$$f(x) = 0 = 3e^x + xe^x - 9.$$

- Second, we find the first derivative to f:

$$\begin{aligned} f'(x) &= 3e^x + (e^x + xe^x) - 0 \\ f'(x) &= 4e^x + xe^x. \end{aligned}$$

- The iteration step (Eq. A.2.2) takes the form

$$x_{n+1} = x_n - \frac{3e^{x_n} + x_n e^{x_n} - 9}{4e^{x_n} + x_n e^{x_n}}.$$

- If we start with a guess of $x_0 = 1$, then the first step brings us to

$$\begin{aligned} x_1 &= x_0 - \frac{3e^{x_0} + x_0 e^{x_0} - 9}{4e^{x_0} + x_n e^{x_0}} \\ x_1 &= 1 - \frac{3e + e - 9}{4e + e} \\ x_1 &= 1 - \frac{4e - 9}{5e} = 0.862183, \end{aligned}$$

  which isn't very far from the accurate solution of 0.849326. In fact, in only takes a couple more step to arrive at the accurate solution, but that's only because I started with a good guess. Table A.1 shows what happens when I start with different guesses.

Table A.1: This table contains a few worked out examples of Newton's method from Example A.2.1. Notice that all guesses, x_0, arrive at the same result. The better guess just takes fewer steps.

$f(x)$	$f'(x)$	x_n
		$x_0 = 1$
1.873127314	13.59140914	0.862182994
0.146904904	11.51523000	0.849425550
0.001124187	11.33942741	0.849326410
0	11.33807149	0.849326404
		$x_0 = 5$
1178.305273	1335.718432	4.117849058
428.2279305	498.6549048	3.259082953
153.8967613	188.9224207	2.444480002
53.74521086	74.26976616	1.720831422
17.38554585	31.97471934	1.177103558
4.554541204	16.79950294	0.905991893
0.664927877	12.13931291	0.851217139
0.021461754	11.36395802	0.849328559
0	11.33810087	0.849326404
		$x_0 = 10$
286335.0553	308370.5211	9.071457757
105052.5408	113764.8426	8.148039429
38525.26312	41990.85863	7.230571571
14119.53827	15509.54990	6.320194522
5170.055703	5734.736777	5.418661304
1890.055866	2124.632808	4.529069531
688.7361317	790.4084239	3.657702133
249.1334070	296.9055541	2.818602279
88.48147441	114.2348921	2.044044937
29.94900647	46.67078670	1.402337167
8.894130266	21.95881900	0.997300345
1.836494564	13.54744786	0.861740158
0.141806903	11.50908345	0.849418855
0.001048270	11.33933584	0.849326409
0	11.33807148	0.849326404

A.3 Orders of Magnitude

There are times when we don't need to know an exact value and sometimes even a slightly approximate value is unnecessary. In those cases, we usually resort to an **order of magnitude** approximation (i.e. all we're concerned with is its power of ten). Unfortunately, this isn't as easy as rounding simple numbers. Simple numbers round by checking the next decimal place then

- rounding up if it's greater than or equal to 5 or
- rounding down if it's less than 5.

For example, 3.4 rounds to 3, but 3.6 rounds to 4.

For an order of magnitude, your first thought might be to just put a number in scientific notation like 4.4×10^4 and rounding the 4.4 to 1 (the nearest power of ten) getting 10^4. If you did this though, you'd be wrong. Formally, an order of magnitude is defined as

$$\log_{10}(\text{number}) \text{ rounded to the nearest integer.} \tag{A.3.1}$$

The consequence is that

$$\log_{10}\left(4.4 \times 10^4\right) = 4.64 \approx 5,$$

so 4.4×10^4 actually rounds up to 10^5. In fact, any front number bigger than $\sqrt{10} \approx 3.162$ will round up. It's a bit strange, but it's the scientific standard, so you should know it.

Appendix B

Useful Formulas

B.1 Single-Variable Calculus

For the following formulas, we have real-valued functions $f(x)$ and $g(x)$ and real-valued constant c.

- *Fundamental Theorem of Calculus (or Inverse Property)*:

$$\int_a^b \frac{d}{dx}(f)\,dx = \int_a^b df = f|_{x=b} - f|_{x=a}$$

- *Chain Rule*:

$$\frac{d}{dx}(f) = \frac{d}{du}(f)\frac{du}{dx}$$

- *Constant Multiple Property*:

$$c\frac{d}{dx}(f) = \frac{d}{dx}(cf)$$

- *Distributive Property*:

$$\frac{d}{dx}(f+g) = \frac{d}{dx}(f) + \frac{d}{dx}(g)$$

- *Product Rule*:

$$\frac{d}{dx}(f*g) = \frac{d}{dx}(f)*g + f*\frac{d}{dx}(g)$$

B.2 Multi-Variable Calculus

For the following formulas, we have vector fields $\vec{A}(q_1, q_2, q_3)$ and $\vec{B}(q_1, q_2, q_3)$, and scalar functions $f(q_1, q_2, q_3)$ and $g(q_1, q_2, q_3)$ given that we're working in the generalized coordinates (q_1, q_2, q_3) with orthonormal unit vectors $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$ and scale factors $\{h_1, h_2, h_3\}$.

- *Path Element*:

$$d\vec{\ell} = h_1\hat{e}_1\, dq_1 + h_2\hat{e}_2\, dq_2 + h_3\hat{e}_3\, dq_3$$

- *Volume Element*:

$$dV = (h_1 dq_1)\,(h_2 dq_2)\,(h_3 dq_3) = h_1 h_2 h_3\, dq_1 dq_2 dq_3$$

- *Fundamental Theorem of Vector Calculus*:

$$\int_a^b \vec{\nabla} f \bullet d\vec{\ell} = \int_a^b df = f|_{x=b} - f|_{x=a}$$

- *Gradient*:

$$\vec{\nabla} f = \sum_{i=1}^{3} \frac{1}{h_i}\frac{\partial f}{\partial q_i}\hat{e}_i = \frac{1}{h_1}\frac{\partial f}{\partial q_1}\hat{e}_1 + \frac{1}{h_2}\frac{\partial f}{\partial q_2}\hat{e}_2 + \frac{1}{h_3}\frac{\partial f}{\partial q_3}\hat{e}_3$$

  written compact and expanded.

- *Divergence*:

$$\begin{aligned}
\vec{\nabla} \bullet \vec{A} &= \frac{1}{h_1 h_2 h_3}\sum_{i=1}^{3}\frac{\partial}{\partial q_i}\left(H_i A_i\right) \\
\vec{\nabla} \bullet \vec{A} &= \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial q_1}\left(h_2 h_3 A_1\right) + \frac{\partial}{\partial q_2}\left(h_3 h_1 A_1\right) + \frac{\partial}{\partial q_3}\left(h_1 h_2 A_1\right)\right]
\end{aligned}$$

  written compact and expanded where $\vec{H} = (h_2 h_3)\,\hat{e}_1 + (h_3 h_1)\,\hat{e}_2 + (h_1 h_2)\,\hat{e}_3$ (the even permutations of the subscripts).

- *Curl*:

$$\begin{aligned}
\vec{\nabla} \times \vec{A} &= \det \begin{bmatrix} \frac{1}{h_2 h_3}\hat{e}_1 & \frac{1}{h_1 h_3}\hat{e}_2 & \frac{1}{h_1 h_2}\hat{e}_3 \\ \frac{\partial}{\partial q_1} & \frac{\partial}{\partial q_2} & \frac{\partial}{\partial q_3} \\ h_1 A_1 & h_2 A_2 & h_3 A_3 \end{bmatrix} \\
\vec{\nabla} \times \vec{A} &= \frac{1}{h_2 h_3}\left[\frac{\partial}{\partial q_2}(h_3 A_3) - \frac{\partial}{\partial q_3}(h_2 A_2)\right]\hat{e}_1 \\
&\quad -\frac{1}{h_1 h_3}\left[\frac{\partial}{\partial q_1}(h_3 A_3) - \frac{\partial}{\partial q_3}(h_1 A_1)\right]\hat{e}_2 \\
&\quad +\frac{1}{h_1 h_2}\left[\frac{\partial}{\partial q_1}(h_2 A_2) - \frac{\partial}{\partial q_2}(h_1 A_1)\right]\hat{e}_3
\end{aligned}$$

  written compact and expanded.

- *Laplacian*:

$$\vec{\nabla}^2 f = \vec{\nabla} \bullet \left(\vec{\nabla} f\right) = \frac{1}{h_1 h_2 h_3}\sum_{i=1}^{3}\frac{\partial}{\partial q_i}\left(H_i \frac{1}{h_i}\frac{\partial f}{\partial q_i}\right)$$

  where $\vec{H} = (h_2 h_3)\,\hat{e}_1 + (h_3 h_1)\,\hat{e}_2 + (h_1 h_2)\,\hat{e}_3$ (the even permutations of the subscripts).

- *Divergence Theorem*:

$$\int \vec{\nabla} \bullet \vec{B}\, dV = \oint_V \vec{B} \bullet d\vec{A}$$

  where $d\vec{A}$ is the area element of the surface enclosing the volume V.

- *Curl Theorem*:

$$\int \left(\vec{\nabla} \times \vec{B}\right) \bullet d\vec{A} = \oint_A \vec{B} \bullet d\vec{\ell}$$

  where $d\vec{\ell}$ is the length element of the path enclosing the area A.

- *Derivative Product Rules*:

$$\begin{aligned}
\vec{\nabla}(fg) &= \left(\vec{\nabla}f\right)g + f\left(\vec{\nabla}g\right) \\
\vec{\nabla}\bullet\left(f\vec{A}\right) &= \vec{A}\bullet\left(\vec{\nabla}f\right) + f\left(\vec{\nabla}\bullet\vec{A}\right) \\
\vec{\nabla}\times\left(f\vec{A}\right) &= -\vec{A}\times\left(\vec{\nabla}f\right) + f\left(\vec{\nabla}\times\vec{A}\right) \\
\vec{\nabla}\bullet\left(\vec{A}\times\vec{B}\right) &= \vec{B}\bullet\left(\vec{\nabla}\times\vec{A}\right) - \vec{A}\bullet\left(\vec{\nabla}\times\vec{B}\right) \\
\vec{\nabla}\left(\vec{A}\bullet\vec{B}\right) &= \vec{A}\times\left(\vec{\nabla}\times\vec{B}\right) + \vec{B}\times\left(\vec{\nabla}\times\vec{A}\right) \\
&\quad + \left(\vec{A}\bullet\vec{\nabla}\right)\vec{B} + \left(\vec{B}\bullet\vec{\nabla}\right)\vec{A} \\
\vec{\nabla}\times\left(\vec{A}\times\vec{B}\right) &= \left(\vec{B}\bullet\vec{\nabla}\right)\vec{A} - \vec{B}\left(\vec{\nabla}\bullet\vec{A}\right) \\
&\quad - \left(\vec{A}\bullet\vec{\nabla}\right)\vec{B} + \vec{A}\left(\vec{\nabla}\bullet\vec{B}\right)
\end{aligned}$$

- *Second Derivative Rules*:

$$\begin{aligned}
\vec{\nabla}\times\left(\vec{\nabla}f\right) &= 0 \\
\vec{\nabla}\bullet\left(\vec{\nabla}\times\vec{A}\right) &= 0 \\
\vec{\nabla}\times\left(\vec{\nabla}\times\vec{A}\right) &= \vec{\nabla}\left(\vec{\nabla}\bullet\vec{A}\right) - \vec{\nabla}^2\vec{A}
\end{aligned}$$

Now if you're looking for a particular coordinate system, just use the following. They are sorted as (q_1, q_2, q_3); $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$; and $\{h_1, h_2, h_3\}$.

- *Cartesian*:

$$(x, y, z)\,;\{\hat{x}, \hat{y}, \hat{z}\}\,;\{1, 1, 1\}$$

- *Cylindrical*:

$$(s, \phi, z)\,;\left\{\hat{s}, \hat{\phi}, \hat{z}\right\}\,;\{1, s, 1\}$$

- *Spherical*:

$$(r, \theta, \phi)\,; \left\{\hat{r}, \hat{\theta}, \hat{\phi}\right\}; \{1, r, r\sin\theta\}$$

- *Bipolar Cylindrical*:

$$(\tau, \sigma, z)\,; \{\hat{\tau}, \hat{\sigma}, \hat{z}\}\,; \left\{\frac{a}{\cosh\tau - \cos\sigma}, \frac{a}{\cosh\tau - \cos\sigma}, 1\right\}$$

- *Elliptic Cylindrical*:

$$(\mu, \nu, z)\,; \{\hat{\mu}, \hat{\nu}, \hat{z}\}\,; \left\{a\sqrt{\sinh^2\mu + \sin^2\nu}, a\sqrt{\sinh^2\mu + \sin^2\nu}, 1\right\}$$

B.3 List of Constants

This is a list of constants used throughout this book. Numbers are consistent with 2014 CODATA recommended values wherever possible and are carried out to four significant figures (unless an *exact* value is available).

Name	Symbol	Value
Gravitational constant	G	$= 6.674 \times 10^{-11}$ Nm^2/kg^2
Earth's surface gravity	g	$= 9.807$ m/s^2 $= 9.807$ N/kg
Mass of the Sun	$M_\odot$	$= 1.989 \times 10^{30}$ kg $= 1477$ m (geometrized)
Mass of the Earth	$M_\oplus$	$= 5.972 \times 10^{24}$ kg $= 4.435$ mm (geometrized)
Coulomb's constant	k_E	$= 8.988 \times 10^{9}$ Nm^2/C^2
Permittivity of free space	ϵ_0	$= 8.854 \times 10^{-12}$ C^2/(Nm^2)
Permeability of free space	μ_0	$= 4\pi \times 10^{-7}$ N/A^2
Speed of light	c	$= 299,792,458$ m/s $= 1$ (relativistic units)
Planck's constant	h	$= 6.626 \times 10^{-34}$ J s $= 4.136 \times 10^{-15}$ eV s
Planck's constant/2π	$\hbar$	$= 1.055 \times 10^{-34}$ J s $= 6.582 \times 10^{-16}$ eV s
Mass of the proton	m_p	$= 1.673 \times 10^{-27}$ kg $= 938.3$ MeV/c^2
Mass of the neutron	m_n	$= 1.675 \times 10^{-27}$ kg $= 939.6$ MeV/c^2
Mass of the electron	m_e	$= 9.109 \times 10^{-31}$ kg $= 0.5110$ MeV/c^2
Elementary charge	e	$= 1.602 \times 10^{-19}$ C
Bohr radius	a_0	$= 5.292 \times 10^{-11}$ m $= 0.05292$ nm $= 52.92$ pm
Boltzmann's constant	k_B	$= 1.381 \times 10^{-23}$ J/K $= 8.617 \times 10^{-5}$ eV/K

Appendix C

Useful Spacetime Geometries

This is a list of all the quantities that are relevant to the spacetime geometries I used in Chapters 7 and 8. All information is given in geometrized units. See Table 8.1 for more details on the units.

C.1 Minkowski Geometry (Cartesian)

This is known as flat spacetime.

- *Line Element*:

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2$$

- *Christoffel Symbols*: $\Gamma^{\delta}_{\mu\nu} = 0$
- *Riemann Curvatures*: $R^{\delta}_{\alpha\mu\nu} = 0$
- *Ricci Curvatures*: $R_{\alpha\nu} = 0$
- *Ricci Curvature Scalar*: $R = 0$
- *Kretschmann Invariant*: $K = 0$

C.2 Minkowski Geometry (Spherical)

This is also known as flat spacetime. Notice, even though there are Christoffel symbols, the curvature tensors are still zero just like in Section C.1.

- *Line Element*:

$$ds^2 = -dt^2 + dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta \, d\phi^2$$

- *Christoffel Symbols* ($\Gamma^\delta_{\mu\nu} = \Gamma^\delta_{\nu\mu}$):

$$\Gamma^r_{\theta\theta} = -r \qquad \Gamma^\theta_{\phi\phi} = -\cos\theta \, \sin\theta$$
$$\Gamma^r_{\phi\phi} = -r\sin^2\theta \qquad \Gamma^\phi_{r\phi} = \frac{1}{r}$$
$$\Gamma^\theta_{r\theta} = \frac{1}{r} \qquad \Gamma^\phi_{\theta\phi} = \cot\theta$$

- *Riemann Curvatures*: $R^\delta_{\alpha\mu\nu} = 0$
- *Ricci Curvatures*: $R_{\alpha\nu} = 0$
- *Ricci Curvature Scalar*: $R = 0$
- *Kretschmann Invariant*: $K = 0$

C.3 Schwarzchild Geometry

This geometry applies to the spacetime *outside* of a spherically symmetric and static source of gravity. Notice it reduces to Section C.2 when $M = 0$.

- *Line Element*:

$$ds^2 = -\left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta \, d\phi^2$$

- *Christoffel Symbols* ($\Gamma^\delta_{\mu\nu} = \Gamma^\delta_{\nu\mu}$):

$$\Gamma^t_{tr} = \frac{M}{r^2}\left(1 - \frac{2M}{r}\right)^{-1} \qquad \Gamma^\theta_{r\theta} = \frac{1}{r}$$
$$\Gamma^r_{tt} = \frac{M}{r^2}\left(1 - \frac{2M}{r}\right) \qquad \Gamma^\theta_{\phi\phi} = -\cos\theta \, \sin\theta$$
$$\Gamma^r_{rr} = -\frac{M}{r^2}\left(1 - \frac{2M}{r}\right)^{-1} \qquad \Gamma^\phi_{r\phi} = \frac{1}{r}$$
$$\Gamma^r_{\theta\theta} = -r\left(1 - \frac{2M}{r}\right) \qquad \Gamma^\phi_{\theta\phi} = \cot\theta$$
$$\Gamma^r_{\phi\phi} = -r\left(1 - \frac{2M}{r}\right)\sin^2\theta$$

- *Riemann Curvatures* ($R^{\delta}_{\alpha\mu\nu} = -R^{\delta}_{\alpha\nu\mu}$):

$$R^{t}_{rtr} = \frac{2M}{r^3}\left(1-\frac{2M}{r}\right)^{-1} \qquad R^{\theta}_{t\theta t} = \frac{M}{r^3}\left(1-\frac{2M}{r}\right)$$
$$R^{t}_{\theta t\theta} = -\frac{M}{r} \qquad R^{\theta}_{r\theta r} = -\frac{M}{r^3}\left(1-\frac{2M}{r}\right)^{-1}$$
$$R^{t}_{\phi t\phi} = -\frac{M}{r}\sin^2\theta \qquad R^{\theta}_{\phi\theta\phi} = \frac{2M}{r}\sin^2\theta$$
$$R^{r}_{trt} = -\frac{2M}{r^3}\left(1-\frac{2M}{r}\right) \qquad R^{\phi}_{t\phi t} = \frac{M}{r^3}\left(1-\frac{2M}{r}\right)$$
$$R^{r}_{\theta r\theta} = -\frac{M}{r} \qquad R^{\phi}_{r\phi r} = -\frac{M}{r^3}\left(1-\frac{2M}{r}\right)^{-1}$$
$$R^{r}_{\phi r\phi} = -\frac{M}{r}\sin^2\theta \qquad R^{\phi}_{\theta\phi\theta} = \frac{2M}{r}$$

- *Ricci Curvatures*: $R_{\alpha\nu} = 0$

- *Ricci Curvature Scalar*: $R = 0$

- *Kretschmann Invariant*: $K = \dfrac{48M^2}{r^6}$

C.4 Eddington-Finkelstein Geometry

This geometry is just a change in variable from Section C.3 that eliminates the singularity at $r = 2M$. It is helpful in predicting the path of particles once they pass the event horizon.

- *Line Element*:

$$ds^2 = -\left(1-\frac{2M}{r}\right)(dt^*)^2 + \frac{4M}{r}dt^*dr + \left(1+\frac{2M}{r}\right)dr^2 + r^2d\theta^2 + r^2\sin^2\theta\, d\phi^2$$

- *Christoffel Symbols* ($\Gamma^{\delta}_{\mu\nu} = \Gamma^{\delta}_{\nu\mu}$):

$$\Gamma^t_{tt} = \frac{2M^2}{r^3} \qquad \Gamma^r_{rr} = -\frac{M}{r^2}\left(1+\frac{2M}{r}\right)$$
$$\Gamma^t_{tr} = \frac{M}{r^2}\left(1+\frac{2M}{r}\right) \qquad \Gamma^r_{\theta\theta} = -r\left(1-\frac{2M}{r}\right)$$
$$\Gamma^t_{rr} = \frac{2M}{r^2}\left(1+\frac{M}{r}\right) \qquad \Gamma^r_{\phi\phi} = -r\left(1-\frac{2M}{r}\right)\sin^2\theta$$
$$\Gamma^t_{\theta\theta} = -2M \qquad \Gamma^\theta_{r\theta} = \frac{1}{r}$$
$$\Gamma^t_{\phi\phi} = -2M\sin^2\theta \qquad \Gamma^\theta_{\phi\phi} = -\cos\theta\,\sin\theta$$
$$\Gamma^r_{tt} = \frac{M}{r^2}\left(1-\frac{2M}{r}\right) \qquad \Gamma^\phi_{r\phi} = \frac{1}{r}$$
$$\Gamma^r_{tr} = -\frac{2M^2}{r^3} \qquad \Gamma^\phi_{\theta\phi} = \cot\theta$$

- *Riemann Curvatures* ($R^{\delta}_{\alpha\mu\nu} = -R^{\delta}_{\alpha\nu\mu}$):

$$R^t_{trt} = -\frac{4M^2}{r^4} \qquad R^\theta_{t\theta r} = -\frac{2M^2}{r^4}$$
$$R^t_{rtr} = \frac{2M}{r^3}\left(1+\frac{2M}{r}\right) \qquad R^\theta_{r\theta t} = -\frac{2M^2}{r^4}$$
$$R^t_{\theta t\theta} = -\frac{M}{r} \qquad R^\theta_{r\theta r} = -\frac{M}{r^3}\left(1+\frac{2M}{r}\right)$$
$$R^t_{\phi t\phi} = -\frac{M}{r}\sin^2\theta \qquad R^\theta_{\phi\theta\phi} = \frac{2M}{r}\sin^2\theta$$
$$R^r_{trt} = -\frac{2M}{r^3}\left(1-\frac{2M}{r}\right) \qquad R^\phi_{t\phi t} = \frac{M}{r^3}\left(1-\frac{2M}{r}\right)$$
$$R^r_{rtr} = -\frac{4M^2}{r^4} \qquad R^\phi_{t\phi r} = -\frac{2M^2}{r^4}$$
$$R^r_{\theta r\theta} = -\frac{M}{r} \qquad R^\phi_{r\phi t} = -\frac{2M^2}{r^4}$$
$$R^r_{\phi r\phi} = -\frac{M}{r}\sin^2\theta \qquad R^\phi_{r\phi r} = -\frac{M}{r^3}\left(1+\frac{2M}{r}\right)$$
$$R^\theta_{t\theta t} = \frac{M}{r^3}\left(1-\frac{2M}{r}\right) \qquad R^\phi_{\theta\phi\theta} = \frac{2M}{r}$$

- *Ricci Curvatures*: $R_{\alpha\nu} = 0$
- *Ricci Curvature Scalar*: $R = 0$

- *Kretschmann Invariant*: $K = \dfrac{48M^2}{r^6}$

C.5 Spherically Symmetric Geometry

This is a generalization of the Schwarzchild geometry from Section C.3 making some of the coefficients arbitrary functions of r. It allows for analysis *inside* a spherically symmetric and static source of gravity. It will reduce to the Schwarzchild geometry outside (i.e. $r > R_{\text{source}}$).

- *Line Element*:

$$ds^2 = -a(r)\,dt^2 + b(r)\,dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2$$

  where a and b are arbitrary functions of radial distance from the center of the source of gravity.

- *Christoffel Symbols* ($\Gamma^\delta_{\mu\nu} = \Gamma^\delta_{\nu\mu}$):

$$\begin{aligned}
\Gamma^t_{tr} &= \frac{1}{2a}\frac{\partial a}{\partial r} & \Gamma^\theta_{r\theta} &= \frac{1}{r} \\
\Gamma^r_{tt} &= \frac{1}{2b}\frac{\partial a}{\partial r} & \Gamma^\theta_{\phi\phi} &= -\cos\theta\,\sin\theta \\
\Gamma^r_{rr} &= \frac{1}{2b}\frac{\partial b}{\partial r} & \Gamma^\phi_{r\phi} &= \frac{1}{r} \\
\Gamma^r_{\theta\theta} &= -\frac{r}{b} & \Gamma^\phi_{\theta\phi} &= \cot\theta \\
\Gamma^r_{\phi\phi} &= -\frac{r}{b}\sin^2\theta
\end{aligned}$$

- *Riemann Curvatures* ($R^\delta_{\alpha\mu\nu} = -R^\delta_{\alpha\nu\mu}$):

$$\begin{aligned}
R^t_{rtr} &= \frac{1}{4a^2}\left(\frac{\partial a}{\partial r}\right)^2 + \frac{1}{4ab}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} - \frac{1}{2a}\frac{\partial^2 a}{\partial r^2} & R^\theta_{t\theta t} &= \frac{1}{2rb}\frac{\partial a}{\partial r} \\
R^t_{\theta t\theta} &= -\frac{r}{2ab}\frac{\partial a}{\partial r} & R^\theta_{r\theta r} &= \frac{1}{2rb}\frac{\partial b}{\partial r} \\
R^t_{\phi t\phi} &= -\left(\frac{r}{2ab}\frac{\partial a}{\partial r}\right)\sin^2\theta & R^\theta_{\phi\theta\phi} &= \left(1 - \frac{1}{b}\right)\sin^2\theta \\
R^r_{trt} &= -\frac{1}{4ab}\left(\frac{\partial a}{\partial r}\right)^2 - \frac{1}{4b^2}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} + \frac{1}{2b}\frac{\partial^2 a}{\partial r^2} & R^\phi_{t\phi t} &= \frac{1}{2rb}\frac{\partial a}{\partial r} \\
R^r_{\theta r\theta} &= \frac{r}{2b^2}\frac{\partial b}{\partial r} & R^\phi_{r\phi r} &= \frac{1}{2rb}\frac{\partial b}{\partial r} \\
R^r_{\phi r\phi} &= \left(\frac{r}{2b^2}\frac{\partial b}{\partial r}\right)\sin^2\theta & R^\phi_{\theta\phi\theta} &= 1 - \frac{1}{b}
\end{aligned}$$

- *Ricci Curvatures* ($R_{\alpha\nu} = R_{\nu\alpha}$):

$$R_{tt} = \frac{1}{rb}\frac{\partial a}{\partial r} - \frac{1}{4ab}\left(\frac{\partial a}{\partial r}\right)^2 - \frac{1}{4b^2}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} + \frac{1}{2b}\frac{\partial^2 a}{\partial r^2}$$
$$R_{rr} = \frac{1}{4a^2}\left(\frac{\partial a}{\partial r}\right)^2 + \frac{1}{rb}\frac{\partial b}{\partial r} + \frac{1}{4ab}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} - \frac{1}{2a}\frac{\partial^2 a}{\partial r^2}$$
$$R_{\theta\theta} = 1 - \frac{1}{b} - \frac{r}{2ab}\frac{\partial a}{\partial r} + \frac{r}{2b^2}\frac{\partial b}{\partial r}$$
$$R_{\phi\phi} = \left(1 - \frac{1}{b} - \frac{r}{2ab}\frac{\partial a}{\partial r} + \frac{r}{2b^2}\frac{\partial b}{\partial r}\right)\sin^2\theta$$

- *Ricci Curvature Scalar*:

$$R = \frac{2}{r^2}\left(1 - \frac{1}{b}\right) - \frac{2}{rab}\frac{\partial a}{\partial r} + \frac{1}{2a^2 b}\left(\frac{\partial a}{\partial r}\right)^2 + \frac{2}{rb^2}\frac{\partial b}{\partial r} + \frac{1}{2ab^2}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r} - \frac{1}{ab}\frac{\partial^2 a}{\partial r^2}$$

- *Kretschmann Invariant*:

$$\begin{aligned} K \;=\; & \frac{4}{r^4} + \frac{4}{r^4 b^2} - \frac{8}{r^4 b} + \frac{2}{r^2 a^2 b^2}\left(\frac{\partial a}{\partial r}\right)^2 + \frac{1}{4a^4 b^2}\left(\frac{\partial a}{\partial r}\right)^4 \\ & + \frac{1}{2a^3 b^3}\left(\frac{\partial a}{\partial r}\right)^3 \frac{\partial b}{\partial r} + \frac{2}{r^2 b^4}\left(\frac{\partial b}{\partial r}\right)^2 + \frac{1}{4a^2 b^4}\left(\frac{\partial a}{\partial r}\right)^2\left(\frac{\partial b}{\partial r}\right)^2 \\ & - \frac{1}{a^3 b^2}\left(\frac{\partial a}{\partial r}\right)^2 \frac{\partial^2 a}{\partial r^2} - \frac{1}{a^2 b^3}\frac{\partial a}{\partial r}\frac{\partial b}{\partial r}\frac{\partial^2 a}{\partial r^2} + \frac{1}{a^2 b^2}\left(\frac{\partial^2 a}{\partial r^2}\right)^2 \end{aligned}$$

C.6 Cosmological Geometry

This is also known as the Friedmann-Lemaître-Robertson-Walker geometry. It is considered the standard model of cosmology by the scientific community. Notice it's still spherically symmetric since the universe has no angular dependence, but the space components do change with time.

- *Line Element*:

$$ds^2 = -dt^2 + [a(t)]^2\left[\frac{1}{1-kr^2}\,dr^2 + r^2 d\theta^2 + r^2\sin^2\theta\, d\phi^2\right]$$

  where k is a constant and $a(t)$ is a function of time.

- *Christoffel Symbols* ($\Gamma^{\delta}_{\mu\nu} = \Gamma^{\delta}_{\nu\mu}$):

$$\Gamma^{t}_{rr} = \frac{a}{1-kr^2}\frac{\partial a}{\partial t} \qquad \Gamma^{r}_{\theta\theta} = -r\left(1-kr^2\right) \qquad \Gamma^{\phi}_{t\phi} = \frac{1}{a}\frac{\partial a}{\partial t}$$

$$\Gamma^{t}_{\theta\theta} = r^2\left(a\frac{\partial a}{\partial t}\right) \qquad \Gamma^{r}_{\phi\phi} = -r\sin^2\theta\,\left(1-kr^2\right) \qquad \Gamma^{\phi}_{r\phi} = \frac{1}{r}$$

$$\Gamma^{t}_{\phi\phi} = r^2\sin^2\theta\,\left(a\frac{\partial a}{\partial t}\right) \qquad \Gamma^{\theta}_{t\theta} = \frac{1}{a}\frac{\partial a}{\partial t} \qquad \Gamma^{\phi}_{\theta\phi} = \cot\theta$$

$$\Gamma^{r}_{tr} = \frac{1}{a}\frac{\partial a}{\partial t} \qquad \Gamma^{\theta}_{r\theta} = \frac{1}{r}$$

$$\Gamma^{r}_{rr} = \frac{kr}{1-kr^2} \qquad \Gamma^{\theta}_{\phi\phi} = -\cos\theta\,\sin\theta$$

- *Riemann Curvatures* ($R^{\delta}_{\alpha\mu\nu} = -R^{\delta}_{\alpha\nu\mu}$):

$$R^{t}_{rtr} = \frac{a}{1-kr^2}\frac{\partial^2 a}{\partial t^2} \qquad R^{\theta}_{t\theta t} = -\frac{1}{a}\frac{\partial^2 a}{\partial t^2}$$

$$R^{t}_{\theta t\theta} = r^2\left(a\frac{\partial^2 a}{\partial t^2}\right) \qquad R^{\theta}_{r\theta r} = \frac{1}{1-kr^2}\left[k+\left(\frac{\partial a}{\partial t}\right)^2\right]$$

$$R^{t}_{\phi t\phi} = r^2\sin^2\theta\,\left(a\frac{\partial^2 a}{\partial t^2}\right) \qquad R^{\theta}_{\phi\theta\phi} = r^2\sin^2\theta\,\left[k+\left(\frac{\partial a}{\partial t}\right)^2\right]$$

$$R^{r}_{trt} = -\frac{1}{a}\frac{\partial^2 a}{\partial t^2} \qquad R^{\phi}_{t\phi t} = -\frac{1}{a}\frac{\partial^2 a}{\partial t^2}$$

$$R^{r}_{\theta r\theta} = r^2\left[k+\left(\frac{\partial a}{\partial t}\right)^2\right] \qquad R^{\phi}_{r\phi r} = \frac{1}{1-kr^2}\left[k+\left(\frac{\partial a}{\partial t}\right)^2\right]$$

$$R^{r}_{\phi r\phi} = r^2\sin^2\theta\,\left[k+\left(\frac{\partial a}{\partial t}\right)^2\right] \qquad R^{\phi}_{\theta\phi\theta} = r^2\left[k+\left(\frac{\partial a}{\partial t}\right)^2\right]$$

- *Ricci Curvatures* ($R_{\alpha\nu} = R_{\nu\alpha}$):

$$R_{tt} = -\frac{3}{a}\frac{\partial^2 a}{\partial t^2}$$

$$R_{rr} = \frac{1}{1-kr^2}\left[2k+2\left(\frac{\partial a}{\partial t}\right)^2+a\frac{\partial^2 a}{\partial t^2}\right]$$

$$R_{\theta\theta} = r^2\left[2k+2\left(\frac{\partial a}{\partial t}\right)^2+a\frac{\partial^2 a}{\partial t^2}\right]$$

$$R_{\phi\phi} = r^2\sin^2\theta\,\left[2k+2\left(\frac{\partial a}{\partial t}\right)^2+a\frac{\partial^2 a}{\partial t^2}\right]$$

- *Ricci Curvature Scalar*:

$$R = \frac{6}{a^2}\left[k + \left(\frac{\partial a}{\partial t}\right)^2 + a\frac{\partial^2 a}{\partial t^2}\right]$$

- *Kretschmann Invariant*:

$$K = \frac{12}{a^4}\left[k^2 + 2k\left(\frac{\partial a}{\partial t}\right)^2 + \left(\frac{\partial a}{\partial t}\right)^4 + a^2\left(\frac{\partial^2 a}{\partial t^2}\right)^2\right]$$

Appendix D

Particle Physics

Beyond the use of quantum mechanics (Chapters 9 and 10), the physics of particles is mostly just lists, tables, and diagrams. I felt it was more fitting to include them in an appendix rather than an actual chapter.

D.1 Categorizing by Spin

There are hundreds of different types of quantum particles, each defined by its inherent or "intrinsic" properties:

- rest mass m_p (or rest energy E_p),
- electric charge q, and
- spin s.

For example, an electron is *defined* by $m = 9.109\times10^{-31}$ kg, $q = 1.602\times10^{-19}$ C, and $s = 1/2$, so all electrons are *absolutely identical*. To make sense of all these different particles, we separate them into two major categories:

1. **Fermions** ($s = \frac{1}{2},\ \frac{3}{2},\ \frac{5}{2},\ \frac{7}{2},\ \ldots$)
2. **Bosons** ($s = 0,\ 1,\ 2,\ 3,\ \ldots$)

Even though they're technically categorized by their spin quantum number s, particles in a category do share similar properties:

1. Identical fermions cannot occupy the same state at the same time. At least one of their quantum numbers must be different.

- For example, as shown in Figure 10.10, you can only put two electrons in one orbital if they have opposite spins (i.e. opposite values of m_s).

2. Identical bosons, on the other hand, can occupy the same state at the same time. In fact, there's no limit to how many you can cram into a single state.

You can find a sample lists of particles in Tables D.1 and D.2.

D.2 Fundamental Particles

It turns out many of these hundreds of different particles are made of only a few particles. They're called fundamental particles because, as far as we know, they're not made of anything else. They're also considered to be the only true *point* particles (i.e. as far as we know, they don't have size) and we categorize them as follows:

- Six **Leptons**
  - Electron, Electron-Neutrino, Muon, Muon-Neutrino, Tauon, and Tauon-Neutrino
- Six **Quarks**
  - Up, Down, Charm, Strange, Top, and Bottom
- Five **Force Carriers**
  - Photon, Gluon, ±W-Bosons, and Z-Boson

You can find the full list (with properties) in Table D.1.

The force carrier particles do exactly what their name suggests. They facilitate one of the four fundamental forces.

1. Strong Nuclear Interaction: Gluon (g)
2. Weak Nuclear Interaction: W and Z Bosons ($W^{\pm}$ and Z^0)
3. Electromagnetism: Photon (γ)

4. Gravity: *Unknown*

We have yet to find a quantum mechanism for gravity. General relativity (Chapter 8) seems to be mostly incompatible quantum mechanics (Chapters 9 and 10), which is a huge problem for our understanding of the universe. Fortunately, gravity is very weak in comparison to the other forces, so we can usually ignore it on the quantum level.

Quarks and leptons bond together using force carriers to form atom-like objects. They're also further separated into **families** (labeled in Table D.1) that share similar properties. For example, up quarks interact with electrons in exactly the same way that charm quarks interact with muons. The three types of neutrinos are a tricky bunch though. You'll notice question marks for their mass in Table D.1. We don't have a good measurement of their masses for two reasons.

1. Their masses are extremely small and they're electrically neutral, so they *barely* ever interact with anything else.

2. They're not stable. They randomly switch back and forth between each other (and between different mass eigenstates).

However, we do know they're non-zero. We also have upper and lower limits for those masses, but those change so often it was silly to include them in something as static as a book.

D.3 Building Larger Particles

All other particles, beyond those in Table D.1, are called **hadrons** and are just combinations of quarks bonded using gluons. I should note that it's impossible for a quark to exist without being bonded to *at least* one other quark, so we've never actually "seen" a quark. We just accept the quark model as scientifically valid because it makes extremely accurate predictions. A sample list with particle properties is given in Table D.2.

We've discovered that quarks have an additional property, like charge, but it's not inherent to the quark-type. Unlike charge (which can only go two ways: positive or negative), this quark property can go *three* different ways. In order for a quark combination to be stable, the property must become "neutral." We see this sort of thing in optics with light colors (see Figure

Table D.1: This is the full list of fundamental particles and their properties. Mass is given in units of MeV/c^2, charge in units of the elementary charge e, and spin in units of $\hbar$.

Name	Symbol	Mass	Charge	Spin	Family
Electron	e^-	0.511	-1	1/2	1
Electron-Neutrino	ν_e	?	0	1/2	1
Muon	μ^-	106	-1	1/2	2
Muon-Neutrino	ν_μ	?	0	1/2	2
Tauon	τ^-	1,777	-1	1/2	3
Tauon-Neutrino	ν_τ	?	0	1/2	3
Up Quark	u	2.3	$+2/3$	1/2	1
Down Quark	d	4.8	$-1/3$	1/2	1
Charm Quark	c	1,275	$+2/3$	1/2	2
Strange Quark	s	95	$-1/3$	1/2	2
Top Quark	t	173,070	$+2/3$	1/2	3
Bottom Quark	b	4,180	$-1/3$	1/2	3
Photon	γ	0	0	1	none
Gluon	g	0	0	1	none
W-Bosons	W^+	80,400	$+1$	1	none
	W^-	80,400	-1	1	none
Z-Boson	Z^0	91,200	0	1	none

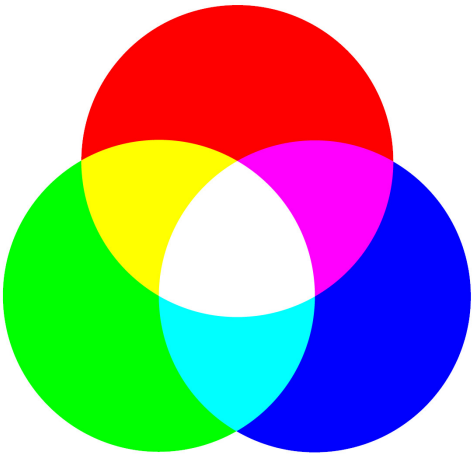


Figure D.1: This chart from optics shows how light colors can add together to make other colors. Red, green, and blue are the primary light colors. Cyan (blue + green), magenta (red + blue), and yellow (red + green) are the secondary light colors. This pattern is used as an *analog* for quantum chromodynamics (Section D.3).

D.1), so we've arbitrarily borrowed the labels: red, green, and blue. As a result, the study of how quarks bond has come to be known as **quantum chromodynamics**. However, quarks don't *actually* have color. It's just an analogy.

Based on Figure D.1, we have a couple ways we can combine quarks using **color charge** and they each have names.

1. **Baryons**: Quarks Triplet
    - One red, one green, and one blue combine to make white (i.e. neutral).
2. **Mesons**: Quark Doublet
    - One red and one anti-red (cyan) combine to make white (i.e. neutral).
    - One green and one anti-green (magenta) combine to make white (i.e. neutral).
    - One blue and one anti-blue (yellow) combine to make white (i.e. neutral).

Technically, the model allows for combinations of more than three quarks, but they're purely hypothetical (i.e. they've never been detected).

There also exists an **anti-particle** (usually signified by a line over the symbol) for *every* particle. When a particle and its anti-particle (e.g. electron and positron) combine, they annihilate each other to generate one or more high energy photons (or Z bosons if the energy is high enough). A hadron's anti-particle is *always* made of the opposite quarks (e.g. uud and $\overline{\text{uud}}$ for the proton and anti-proton, respectively). Sometimes the anti-particle is itself (e.g. $\text{s}\overline{\text{s}}$ for the phi-meson), but it still technically has one. Even weirder, the neutral pion is in a superposition of two quark doublets:

$$\pi^0 \to \frac{\text{u}\overline{\text{u}} - \text{d}\overline{\text{d}}}{\sqrt{2}},$$

meaning it has an equal probability of being $\text{u}\overline{\text{u}}$ or $\text{d}\overline{\text{d}}$ when measured. Since there are six quarks and six anti-quarks, there are around $12^3 = 1{,}728$ potential baryons and around $12^2 = 144$ potential mesons.

D.4 Feynman Diagrams

Quantum field theory can get rather complex and the calculations can sometimes seem impossible. So in 1948, Richard Feynman proposed an alternative method. He took all the particles, their motions, and their interactions and he gave them all symbols for visual analysis in a spacetime diagram (see Section 7.2 for more details). He *literally* turned a nasty field calculation into a picture! When we collide particles in an accelerator, we know what particles we started with and we detect what particles were created in the end. What **Feynman diagrams** do is they give us a simple way of figuring out the most probable interactions in between (and how probable each of them is) without having to do much math.

Feynman diagrams are drawn using a set of consistent rules about particles, interactions, and time. Those rules are as follows.

1. Interactions are drawn as points (since they're *events* in spacetime).

2. Leptons, quarks, and hadrons are all drawn as straight solid lines with arrows (i.e. time-like paths).

Table D.2: This is a sample list of particles and their intrinsic properties. It is *by no means* a complete list. Mass is given in units of $\mathrm{MeV/c^2}$, charge in units of the elementary charge e, and spin in units of $\hbar$. Anti-quarks are signified by a line over the $\overline{\text{symbol}}$.

Name	Symbol	Mass	Charge	Spin	Quarks
Proton	p^+	938.3	+1	1/2	uud
Neutron	n^0	939.6	0	1/2	udd
Deltas	Δ^{++}	1,232	+2	3/2	uuu
	Δ^-	1,232	−1	3/2	ddd
Lambdas	Λ_c^+	2,286	+1	1/2	udc
	Λ_s^0	1,116	0	1/2	uds
Xis	Ξ^0	1,315	0	1/2	uss
	Ξ^-	1,322	−1	1/2	dss
Omega	Ω^-	1,672	−1	3/2	sss
Pions	π^+	139.6	+1	0	$\mathrm{u\overline{d}}$
	π^-	139.6	−1	0	$\mathrm{\overline{u}d}$
	π^0	135.0	0	0	$(\mathrm{u\overline{u}}-\mathrm{d\overline{d}})/\sqrt{2}$
Kaons	K^+	493.7	+1	1	$\mathrm{u\overline{s}}$
	K^-	493.7	−1	1	$\mathrm{\overline{u}s}$
	K^0	497.6	0	1	$\mathrm{d\overline{s}}$
Phi	φ^0	1,019	0	0	$\mathrm{s\overline{s}}$
Upsilon	Υ^0	9,460	0	0	$\mathrm{b\overline{b}}$
J/Psi	J/ψ^0	3,097	0	0	$\mathrm{c\overline{c}}$

- Arrows point toward an interaction for incoming *regular* particles and away from an interaction for outgoing *regular* particles (as you'd expect).
- Arrows point away from an interaction for incoming *anti*-particles and toward an interaction for outgoing *anti*-particles (as if they're regular particles traveling back in time).

3. Photons, W bosons, and Z bosons are all drawn as wavy lines.
4. Gluons are drawn as spirals.

Each item in the diagrams represents a factor in the calculation. Some examples are shown in Figures D.2 and D.3.

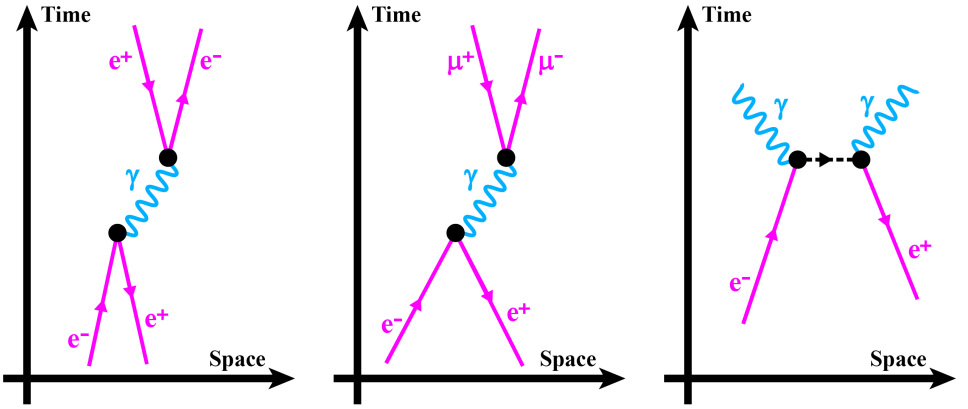


Figure D.2: These are three possible results of an electron-positron annihilation. Far left: The annihilation forms a photon, but then the photon recreates the electron-positron pair. Middle: The electron and positron have enough kinetic energy to generate a slower muon-antimuon pair. Far right: The electron-positron pair just creates two photons that move away in opposite directions (via a *virtual* fermion).

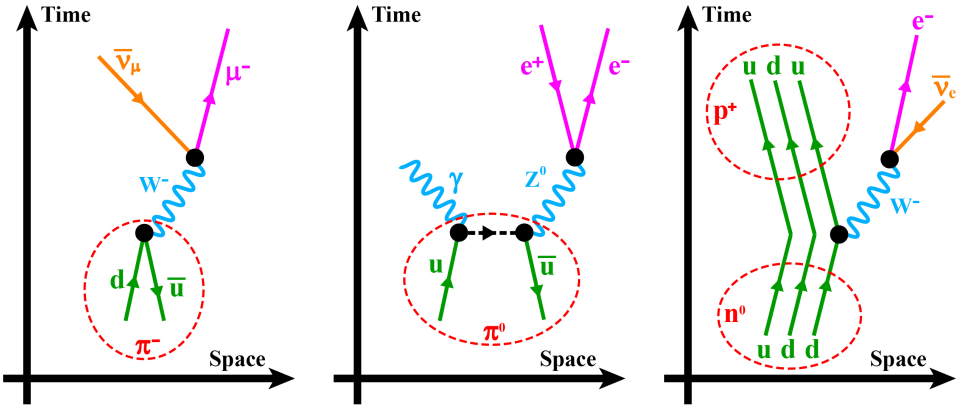


Figure D.3: These are some complex examples of Feynman diagrams. Far left: Decay of a negative pion, π^-. Middle: Decay of a neutral pion, π^0. Far right: Neutron decays into a proton (i.e. negative beta decay).

Index