Intuitive Explanations is a collection of clear, easy-to-understand, intuitive explanations of many of the more complex concepts in calculus. In this document, an “intuitive” explanation is not one without any mathematical symbols. Rather, it is one in which each definition, concept, and part of an equation can be understood in a way that is more than “a piece of mathematics”. For instance, some textbooks simply define mathematical terms, whereas Intuitive Explanations explains what idea each term is meant to represent and why the definition has to be the way it is in order for that idea to be captured. As another example, the proofs in Intuitive Explanations avoid steps with no clear motivation (like adding and subtracting the same quantity), so that they seem natural and not magical. See the table of contents for information on what topics are covered. Before each section, there is a note clarifying what prerequisite knowledge about calculus is required to understand the material presented. Section 1 contains a number of links to other resources useful for learning calculus around the Web. I originally wrote Intuitive Explanations in high school, during the summer between AP Calculus BC and Calculus 3, and the topics covered here are, generally speaking, the topics that are covered poorly in the textbook we used (Calculus by Ron Larson and Bruce Edwards, tenth edition).
# About *Intuitive Explanations*

Since I originally wrote *Intuitive Explanations* while reading my high school Calculus textbook, I happen to have a handy table correlating sections from that textbook to sections in *Intuitive Explanations*:

<table>
<thead>
<tr>
<th><em>Calculus</em> by Larson and Edwards</th>
<th>Topic</th>
<th><em>Intuitive Explanations</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1.3</td>
<td>The formal definition of limit</td>
<td>Section 2</td>
</tr>
<tr>
<td>Chapter 2.3</td>
<td>The product rule</td>
<td>Section 3.1</td>
</tr>
<tr>
<td></td>
<td>The quotient rule</td>
<td>Section 3.2</td>
</tr>
<tr>
<td>Chapter 4.2, 4.3</td>
<td>The formal definition of the integral</td>
<td>Section 4.1</td>
</tr>
<tr>
<td>Chapter 4.4</td>
<td>The first fundamental theorem of calculus</td>
<td>Section 4.2</td>
</tr>
<tr>
<td></td>
<td>The second fundamental theorem of calculus</td>
<td>Section 4.3</td>
</tr>
<tr>
<td>Chapter 9.1</td>
<td>Bounded monotonic sequences</td>
<td>Section 5.1</td>
</tr>
<tr>
<td>Chapter 9.3</td>
<td>The integral test and remainder formula</td>
<td>Section 5.2</td>
</tr>
<tr>
<td>Chapter 9.4</td>
<td>The limit comparison test</td>
<td>Section 5.3</td>
</tr>
<tr>
<td>Chapter 9.5</td>
<td>The alternating series test and remainder formula</td>
<td>Section 5.4</td>
</tr>
<tr>
<td>Chapter 9.6</td>
<td>The ratio and root tests</td>
<td>Section 5.5</td>
</tr>
<tr>
<td>Chapter 13.6</td>
<td>The gradient</td>
<td>Section 6.1</td>
</tr>
<tr>
<td>Chapter 15.1, 15.3</td>
<td>Conservative vector fields</td>
<td>Section 6.8</td>
</tr>
<tr>
<td>Chapter 15.1</td>
<td>The curl</td>
<td>Section 6.5</td>
</tr>
<tr>
<td></td>
<td>The divergence</td>
<td>Section 6.3</td>
</tr>
<tr>
<td>Chapter 15.2</td>
<td>The gradient theorem</td>
<td>Section 6.2</td>
</tr>
<tr>
<td>Chapter 15.4</td>
<td>Green’s theorem</td>
<td>Section 6.6</td>
</tr>
<tr>
<td>Chapter 15.7</td>
<td>The divergence theorem</td>
<td>Section 6.4</td>
</tr>
<tr>
<td>Chapter 15.8</td>
<td>Stokes’ theorem</td>
<td>Section 6.7</td>
</tr>
</tbody>
</table>

Section numbers are hyperlinked: you can click on a number to jump to that section in this document.

Additionally, following is a list of some places on the Web you can find more information and explanations of things in Calculus; the text typeset in bold is hyperlinked. The list is roughly ordered from most to least useful, in my opinion:

**The Feynman Technique** has nothing *per se* to do with Calculus, but it can be used to great effect to help learn Calculus (or anything else).

**Math Insight** is a website that covers many of the concepts in the third semester of Calculus from an intuitive standpoint. It includes many very helpful explanations and interactive visualizations.

**BetterExplained** is a website with an even greater emphasis on intuition-first learning than this document. It covers topics in everything from Arithmetic to Linear Algebra.
Paul’s Online Math Notes can essentially be used as a very easy-to-follow textbook for the first three semesters of Calculus as well as Differential Equations.

This fully geometric proof of the derivatives of \( \sin \) and \( \cos \) is nothing short of beautiful.

These notes about Lagrange multipliers are excellent and provide a concrete, graphical, intuitive explanation of Lagrange multipliers that I have not seen anything else come close to.

This Math StackExchange thread contains several very appealing explanations of why the determinant has the properties it does.

This Math StackExchange thread contains a concrete, intuitive example that instantly demystified the Extended Mean Value Theorem (used in the proof of l’Hôpital’s rule) for me.

This article on Riemann’s Rearrangement Theorem not only provides an excellent proof of said astonishing theorem but also grants much deeper insight into the fundamental natures of absolutely and conditionally convergent series.

The Interactive Gallery of Quadric Surfaces has a variety of useful visualizations, insights, and tips relating to quadric surfaces.

Math StackExchange is always a useful resource, although whether any given answer will be useful to you is more or less up to the flip of a coin.

Wolfram|Alpha is a “computational knowledge engine” with breathtaking breadth and depth. It can solve the vast majority of elementary, computational math problems, and will sometimes provide step-by-step solutions if you buy a subscription for a few dollars.

These notes about curvature contain a number of useful facts and proofs that I was unable to find elsewhere.

These notes about improper double integrals contain some formal definitions and theorems that I was unable to find elsewhere.

These notes about the second partials test contains some details of its generalization to functions of \( n \) variables that I was unable to find elsewhere.

2 The Formal Definition of Limit

Prerequisite Knowledge. Understand generally what a limit is, and be able to evaluate limits graphically and numerically.

What, exactly, does it mean for a function to “have a limit” at a certain point? You have probably learned that the limit of a function at a point is related to the values of the function near that point, but that these two things are not quite the same. For instance, you are surely aware that the first two functions depicted in figure 2.1 both have the limit \( L \) at \( x = c \), while the third function does not have any limit at \( x = c \).

But what about the function

\[
f(x) = \begin{cases} 
  x & \text{if } x \text{ is rational} \\
  0 & \text{if } x \text{ is irrational}
\end{cases}
\]
\[
\lim_{x \to c} f(x) = L = f(c)
\]

\[
\lim_{x \to c} f(x) = L \neq f(c)
\]

\[
\lim_{x \to c} f(x) \text{ does not exist}
\]

Figure 2.1: Functions with different behaviors at \( x = c \) and their limits at that point.

It would be very difficult to graph this function. Does it have a limit anywhere? If so, where? It is not easy to tell without a precise definition of limit. In this section, we will develop just such a definition. We will start with a simple, straightforward definition and gradually improve its weaknesses until we reach a more complex but completely watertight definition that can be applied to any function whatsoever.

You might wonder what the point of all this is. And indeed, it is certainly possible to understand almost everything in calculus quite well without knowledge of the formal definition of limit. But on the other hand, literally everything in calculus is based on the formal definition of limit – and so you will gain a much deeper understanding of how calculus works by taking the time to study limits in detail.

2.1 First Attempts

Think about how you determine limits from a graph. The main idea is to figure out what the value of the function is “heading towards” as \( x \) approaches \( c \) without actually using the value of the function at \( x = c \).

For example, take the second graph pictured in figure 2.1. We know that the limit ought to be \( L \) at \( x = c \), but why? One answer is, “\( f(x) \) becomes very close to \( L \) near (but not at) \( x = c \).” This leads to a definition of what it means for the statement

\[
\lim_{x \to c} f(x) = L
\]

to be true.

**Definition.** \( \lim_{x \to c} f(x) = L \) if \( f(x) \) becomes very close to \( L \) as \( x \) approaches (but does not reach) \( c \).

Unfortunately, this definition is no good in a proof. What does “very close” mean? Reasonable people could disagree over whether \( 1.001 \) is “very close” to 1. Let’s try defining “very close” – for instance, “differing by less than \( 0.01 \)” might seem reasonable.

**Definition.** \( \lim_{x \to c} f(x) = L \) if \( f(x) \) comes within \( 0.01 \) of \( L \) as \( x \) approaches (but does not reach) \( c \).

Here, as well as later, “\( x \) is within 0.01 of \( y \)” means that the value of \( x \) and the value of \( y \) differ by less than 0.01: that the distance between \( x \) and \( y \) on a number line is less than 0.01. In symbols, \( |x - y| < 0.01 \).
Figure 2.2: According to our second definition, \( \lim_{x \to 1} x = 1.001 \). (Here, \( c = 1 \), \( L = 1.001 \), and \( f(x) = x \).) This is because near \( x = 1 \), the difference between \( f(x) \) and \( L = 1.001 \) is approximately 0.001. And according to our definition, all \( f(x) \) has to do for its limit to exist is come within 0.01 of \( L = 1.001 \) near \( x = c = 1 \). In fact, you can use this same reasoning to show that \( \lim_{x \to 1} x \) is equal to an infinite number of different values, which is clearly ridiculous.

This definition probably seems a little off to you: after all, we do not want a function \( f(x) \) that comes within 0.01 of \( L \), but just barely misses. We want a function that actually “hits” the point \((c, L)\), like the second function shown in figure 2.1. A near miss that would satisfy this definition is shown in figure 2.2.

2.2 Arbitrary Closeness

You can probably see that no matter what numerical requirement we put on how close \( f(x) \) must be to \( L \), it will still be possible for \( f(x) \) to satisfy that requirement, but still “just miss” \((c, L)\). So how can we separate functions like the first two in figure 2.1 from the one pictured in figure 2.2?

The key principle is that functions that “pass through” their limiting point \((c, L)\) will be able to satisfy any requirement on how close \( f(x) \) must get to \( L \). For instance, if you zoomed in on the second graph in figure 2.1, you would see that \( f(x) \) comes within 0.01 of \( L \). But if you zoomed in farther, you would also see that it comes within 0.0001 of \( L \); if you zoomed in unimaginably far, you would see \( f(x) \) come within even \( 10^{-1000} \) of \( L \). This type of zoom is depicted in figure 2.3.

Keeping in mind the property of being able to satisfy any requirement on how close \( f(x) \) must get to \( L \), we can write a new definition.

**Definition.** \( \lim_{x \to c} f(x) = L \) if \( f(x) \) can satisfy any requirement on how close it must be to \( L \) (once \( x \) gets close enough to \( c \)).

Let’s rephrase a few parts of this definition to make everything more explicit:

**Definition.** \( \lim_{x \to c} f(x) = L \) if: For any requirement that \( f(x) \) be within a certain distance of \( L \), that requirement is satisfied when \( x \) is sufficiently close to \( c \).

**Definition.** \( \lim_{x \to c} f(x) = L \) if: For any requirement that \( f(x) \) be within a certain distance of \( L \), there is a second requirement that \( x \) be within a certain distance of \( c \) such that the first requirement is satisfied whenever the second requirement is satisfied.
Figure 2.3: A small portion of the second graph in figure 2.1, magnified so much that the graph appears as a straight line. The symbol $\epsilon$ represents a very small number, such as $10^{-1000}$. You can see from the graph that when $x$ is close enough to $c$, $f(x)$ will come within $\epsilon$ of $L$ (all of the values between $L - \epsilon$ and $L + \epsilon$ are within $\epsilon$ of $L$). Notice that when you zoom in even more, the graph will look exactly the same, but $\epsilon$ will become an even smaller number. This shows that no matter how small of a number $\epsilon$ you can think of, $f(x)$ will eventually get within that distance of $L$ (once $x$ is close enough to $c$).

Figure 2.4: A visual representation of the fact that $\lim_{x \to 0.7} x^2 = 0.49$. In this figure (to scale), $f(x) = x^2$, $c = 0.7$, $\delta = 0.075$, $L = 0.49$, and $\epsilon = 0.15$, but many other values would produce a situation similar to the one shown.

Make sure you understand this last definition, because it is essentially the same as the final definition – just in words instead of symbols. An example, for the function $f(x) = x^2$, is pictured in figure 2.4.

The first requirement in our definition is that $f(x)$ be within $\epsilon = 0.15$ of $L = 0.49$, which is equivalent to $f(x)$ being within the blue band. The second requirement is that $x$ be within $\delta = 0.075$ of $c = 0.7$, which is equivalent to $x$ being within the red band. You can see in figure 2.4 that every point on $y = f(x)$ that is contained in the red band is also contained in the blue band. (This is a consequence of the olive-colored projection of the red band being entirely contained within the blue
Therefore, the first requirement is satisfied whenever the second requirement is satisfied. In order for the limit \( \lim_{x \to 0.7} x^2 = 0.49 \) to exist according to our definition, it must always be possible to draw a red band whose projection lies entirely within the blue band, no matter how thin the blue band is made.

2.3 Using Symbols

If you understand our last definition, it is a relatively simple matter to translate its words into symbols. (Of course, what you should remember are the words, not the symbols – but on the other hand, it is easier to do proofs with symbols than with words.)

Firstly, a requirement that \( x \) be within a certain distance of \( y \) can be described with a single number: the distance (call it \( D \)). This requirement is satisfied when \( |x - y| < D \). So, the first requirement in the definition – that \( f(x) \) be within a certain distance of \( L \) – can be described with a single number, which we will call \( \epsilon \). This requirement is satisfied when \( |f(x) - L| < \epsilon \). Analogously, the second requirement – that \( x \) be within a certain distance of \( c \) – can be described with another number, which we will call \( \delta \). This requirement is satisfied when \( |x - c| < \delta \).

We can now make some substitutions in our last definition:

\[
\lim_{x \to c} f(x) = L \text{ if:} \\
\text{For any requirement that } f(x) \text{ be within a certain distance of } L, \\
\text{there is a second requirement that } x \text{ be within a certain distance of } c \text{ such that the first requirement is satisfied whenever the second requirement is satisfied.}
\]

This gives us a concise and very precise final definition:

\[
\text{Definition. } \lim_{x \to c} f(x) = L \text{ if for every positive number } \epsilon \text{ there is at least one positive number } \delta \text{ such that } |f(x) - L| < \epsilon \text{ whenever } |x - c| < \delta.
\]

By the way, if you’re wondering where the letters \( \epsilon \) and \( \delta \) came from, \( \epsilon \) is the lowercase Greek letter “epsilon” and originally stood for “error”, while \( \delta \) is the lowercase Greek letter “delta” (uppercase \( \Delta \)), which stands for change.

2.4 Examples

2.4.1 Linear Function

\[
\lim_{x \to 2} 3x = 6
\]

If we plug in \( f(x) = 3x, \ c = 2, \) and \( L = 6 \) to our definition, then we get:

\[
\text{Definition. } \lim_{x \to 2} 3x = 6 \text{ if for every positive number } \epsilon \text{ there is at least one positive number } \delta \text{ such that } |3x - 6| < \epsilon \text{ whenever } |x - 2| < \delta.
\]

Figure 2.5 illustrates the situation for this limit.
Figure 2.5: Values of $\epsilon$ and $\delta$ that satisfy our definition of limit.

Our definition asserts that for every number $\epsilon$, there is a number $\delta$ that meets a certain specific condition (if the limit exists, which $\lim_{x \to 2} 3x = 6$ does). Pictured in figure 2.5 is an arbitrary value of $\epsilon$ and a value of $\delta$ that meets the condition of the definition, which is that $|3x - 6| < \epsilon$ whenever $|x - 2| < \delta$. You can see in the figure that if $|x - 2| < \delta$ (meaning that $x$ is within the red band), then $|3x - 6| < \epsilon$ (meaning that $f(x) = 3x$ is within the blue band). Just as before, this is a consequence of the projection of the red band being entirely within the blue band. You can also see that if the red band were made narrower, then the same property would hold true: Therefore, there are many values of $\delta$ that are valid for any given $\epsilon$, not just one. Finally, note that no matter how small the blue band were to be made, the red band could be made small enough so that its projection laid once more entirely within the blue band. This shows that our definition holds true, and so that $\lim_{x \to 2} 3x = 6$ exists.

Now that you understand what is going on in the formal definition of limit, we are ready to use it to formally prove an actual limit – in this case, $\lim_{x \to 2} 3x = 6$. The math is very easy – the hard part is keeping track of when each statement and inequality is true.

According to our definition, we will have to prove a certain statement “for every positive number $\epsilon$. Let’s start by proving that statement (“there is at least one positive number $\delta$ such that $|3x - 6| < \epsilon$ whenever $|x - 2| < \delta$”) for a single $\epsilon$. If we let $\epsilon = 1/10$, then we will have to prove that “there is at least one positive number $\delta$ such that $|3x - 6| < 1/10$ whenever $|x - 2| < \delta$.

The easiest way to prove that there is a number with a certain property (say, that it can be divided evenly by 6 positive integers) is to come up with such a number (12) and then show that the number has the property it is supposed to (1, 2, 3, 4, 6, 12). We will do exactly the same thing in order to prove that there is a number $\delta$ with the property in our definition.

We can start by noting that $|3x - 6| < 1/10$ is equivalent to $|x - 2| < 1/30$ (by dividing both sides by 3). Since this expression is quite similar to $|x - 2| < \delta$, we might try letting $\delta = 1/30$. In that case, whenever $|x - 2| < \delta$, it is also true that $|x - 2| < 1/30$, and so (as we said earlier)
|3x − 6| < 1/10, which is what we wanted to prove. Therefore, the statement “there is at least one positive number δ such that |3x − 6| < ϵ whenever |x − 2| < δ” is true when ϵ = 1/10.

What about when ϵ is some other value? If you understood the above reasoning, extending it to any ϵ is quite easy: just replace 1/10 with ϵ, and replace 1/30 with ϵ/3. Since all we need to do is show that there is an appropriate value of δ for each possible value of ϵ, it is perfectly reasonable to need to know what ϵ is in order to figure out what δ ought to be. (On the other hand, you can’t use x in your expression for δ. If you could, you could just say “δ = twice the distance between x and c”, and x would always be within δ of c no matter what limit you were trying to compute!)

That’s it! A typical textbook proof that \( \lim_{x \to 2} 3x = 6 \) might go as follows, albeit probably without colors:

Suppose we are given ϵ > 0. Let δ = ϵ/3. If |x − 2| < δ, then |x − 2| < ϵ/3 and so |3x − 6| < ϵ. Therefore the limit exists.

Even if this proof is a little on the terse side, now that you know how to do the proof yourself you ought to be able to understand it.

As an exercise, prove that \( \lim_{x \to −3} (−5x) = 15 \). If you run into trouble, refer back to the proof that \( \lim_{x \to 2} 3x = 6 \).

2.4.2 Limit Rules

\[
\lim_{x \to c} kx = kc
\]

There is really nothing new about this limit, since the last limit is just a special case of it, with \( c = 2 \) and \( k = 3 \). However, it is important to be able to prove limits that have variables in them – you just have to be able to keep track of which variables are changing, and when they do. According to our definition:

**Definition.** \( \lim_{x \to c} kx = kc \) if for every positive number ϵ there is at least one positive number δ such that |kx − kc| < ϵ whenever |x − c| < δ.

Here is a short proof, based on our previous one:

Suppose we are given ϵ > 0. Let δ = ϵ/|k|. If |x − c| < δ, then |x − c| < ϵ/|k| and so |kx − kc| < ϵ. Therefore the limit exists.

We use |k| instead of just k because k might be negative.

Really, the best way to thoroughly understand the formal definition of limit (once you have studied its basic principles) is to apply it yourself. Try to write proofs for the following limits:

- \( \lim_{x \to c} k = k \)
- \( \lim_{x \to c} (x + 1) = c + 1 \)
- \( \lim_{x \to c} (kx − x) = kc − c \)

The best way to start is to rewrite the formal definition, plugging in the values for \( f(x) \) and \( L \).

2.4.3 Quadratic Function
Finding the limit of a quadratic function is surprisingly difficult, but it is a very good exercise in proof-writing. We will assume that \( c \) is positive because it makes the proof less tedious. The definition says:

**Definition.** \( \lim_{x \to c} x^2 = c^2 \) if for every positive number \( \epsilon \) there is at least one positive number \( \delta \) such that \( |x^2 - c^2| < \epsilon \) whenever \( |x - c| < \delta \).

The first thing to do is to try to get the \( \epsilon \) expression to look more similar to the \( \delta \) expression. Recalling the properties that \( a^2 - b^2 = (a+b)(a-b) \) and that \( |ab| = |a||b| \), we see that \( |x^2 - c^2| < \epsilon \) is equivalent to \( |x + c||x - c| < \epsilon \). The only difference now is the extra \( |x + c| \) term in the \( \epsilon \) expression. Your first thought might be to let \( \delta = \epsilon/|x + c| \), but as we discussed earlier we cannot use \( x \) in our expression for \( \delta \). So what can we do?

In many of the more complex \( \epsilon-\delta \) proofs (such as this one), some of the steps you take will not be symmetric. For instance, a symmetric step is from \( 2x < 4 \) to \( x < 2 \), while an asymmetric step is from \( x < 2 \) to \( x < 3 \). Both are perfectly valid, but you can only reverse symmetric steps (\( x < 2 \) implies \( 2x < 4 \), but \( x < 3 \) does not imply \( x < 2 \)). This idea will come into play shortly.

First, however, let’s examine what possible magnitudes \( |x + c| \) can have. Since we are trying to prove that \( |x - c| < \delta \) implies \( |x + c||x - c| < \epsilon \), we can assume that \( |x - c| < \delta \) is true. This means that \( x \) is between \( c - \delta \) and \( c + \delta \). Therefore, \( x + c \) must be between \( 2c - \delta \) and \( 2c + \delta \). Since \( c \) and \( \delta \) are positive, \( -2c - \delta \) is clearly less than \( 2c - \delta \). Thus, \( x + c \) must also be between \( -2c - \delta \) and \( 2c + \delta \), which means \( |x + c| \) cannot be greater than \( 2c + \delta \) if \( |x - c| < \delta \).

So what was the point of all that? Well, remember how we cannot simply let \( \delta = \epsilon/|x + c| \) because \( x \) cannot be present in our expression for \( \delta \)? The solution is to use the inequality we found above to transform the expression \( |x + c| \) into \( 2c + \delta \), which does not have an \( x \) in it. To do so, we can use the simple principle that if \( a \leq b \) and \( bx < c \), then \( ax < c \). We know that \( |x + c| < 2c + \delta \), so if \( (2c + \delta)|x - c| < \epsilon \), then \( |x + c||x - c| < \epsilon \).

However, we can’t exactly let \( \delta = \epsilon/(2c + \delta) \), as we would be led in endless circles if we ever tried to evaluate that expression. To resolve the problem, we can use another inequality. The key idea is to decide now that we will not let \( \delta \) ever be more than \( c \). We do not know what our expression for \( \delta \) will be yet, but we’ll make sure to write it so that \( \delta \leq c \). This means that we have the inequality \( 2c + \delta \leq 3c \). Now, all we have to do is use the same principle as we used with \( |x + c| < 2c + \delta \), and we can find that if \( 3c|x - c| < \epsilon \), then \( (2c + \delta)|x - c| < \epsilon \).

So what we have found is that if \( 3c|x - c| < \epsilon \), then \( |x + c||x - c| < \epsilon \). This is almost what we want! It is now quite natural to let \( \delta = \epsilon/3c \), which would be correct if it were not for the fact that we also said earlier that we would make \( \delta \leq c \). It’s relatively simple to satisfy this requirement, though – we can just let \( \delta \) be whichever of \( \epsilon/3c \) or \( c \) is smaller. This way, we know that \( \delta \) cannot be greater than \( c \).

We have finally obtained a formula for \( \delta \)! If you’d like to write it out symbolically, \( \delta = \min(\epsilon/3c,c) \) would do nicely. We can now write out the actual proof that \( \lim_{x \to c} x^2 = c^2 \) where \( c > 0 \):
Suppose we are given $\epsilon > 0$. Let $\delta$ be whichever of $\epsilon/3c$ or $c$ is smaller. Now suppose that $|x - c| < \delta$. Since $\delta \leq \epsilon/3c$, we know that $|x - c| < \epsilon/3c$, and so $3c|x - c| < \epsilon$. We also have that $\delta \leq c$, which means that $2c + \delta < 3c$ and so $(2c + \delta)|x - c| < \epsilon$. Additionally, by our reasoning earlier, $|x - c| < \delta$ implies that $|x + c| < 2c + \delta$. Consequently $|x + c||x - c| < \epsilon$ and $|x^2 - c^2| < \epsilon$. Therefore the limit exists.

Figure 2.6 illustrates that the values of $\delta$ given by our formula $\delta = \min(\epsilon/3c, c)$ really do work.

![Figure 2.6: Various values of $\epsilon$ and the corresponding values of $\delta$ for the function $f(x) = x^2$.](image)

Note, by the way, that $c$ must be included along with $\epsilon/3c$ in the formula for $\delta$ because it keeps problems from cropping up for large values of $\epsilon$. For instance, if $c = 1$, $L = 1$, and $\epsilon = 30$, then the simple formula $\delta = \epsilon/3c$ would give us $\delta = 10$. Unfortunately, $x = 10$ is within $\delta = 10$ of $c = 1$, but $f(x) = 100$ is definitely not within $\epsilon = 30$ of $L = 1$. On the other hand, the more complex formula $\delta = \min(\epsilon/3c, c)$ gives us the much smaller $\delta = 1$. All is well.

It might seem silly that we have to worry about large values of $\epsilon$ (which we normally think of as a small number), but such is the price of rigor.

### 2.4.4 Disproving Limits

$$\lim_{x \to 0} \frac{1}{x} \text{ does not exist}$$

You are doubtless already familiar with limits that do not exist, such as the one above (pictured in figure 2.7). It is also possible to use the formal definition of limit to prove this fact.

Of course, first we need to know what it means for a limit not to exist. Here is the definition:

**Definition.** $\lim_{x \to c} f(x)$ does not exist if the statement $\lim_{x \to c} f(x) = L$ is not true for any number $L$.

In other words, if the limit does not have any value, then it does not exist. We can expand this as:
Definition. \( \lim_{x \to c} f(x) \) does not exist if for every number \( L \) the following is false: for every positive number \( \epsilon \) there is at least one positive number \( \delta \) such that \( |f(x) - L| < \epsilon \) whenever \( |x - c| < \delta \).

We can rephrase the above definition using De Morgan’s laws. You may think you have not heard of these, but in fact you use them all the time. For instance, one of De Morgan’s laws states that if it is false that “all ravens are black”, then it is true that “at least one raven is non-black”. Conversely, if it is false that “at least one raven is yellow”, then it is true that “all ravens are non-yellow”.

You should be able to convince yourself that the above definition and the below definition are equivalent.

Definition. \( \lim_{x \to c} f(x) \) does not exist if for every number \( L \) there is at least one positive number \( \epsilon \) for which there is no positive number \( \delta \) such that \( |f(x) - L| < \epsilon \) whenever \( |x - c| < \delta \).

Intuitively, why does \( \lim_{x \to 0} \frac{1}{x} \) not exist? As \( x \) gets closer to \( c = 0 \) from the right, \( f(x) = \frac{1}{x} \) will get larger and larger. In fact, no matter what number \( L \) you pick, \( f(x) \) will grow larger than \( L \), then larger than \( 2L \), then larger than \( 1000L \) as \( x \) continues to approach \( c = 0 \). Therefore, \( f(x) \) does not get closer and closer to any single number as \( x \) approaches \( c \), and consequently the limit does not exist.

The above definition, written for our particular limit, is as follows:

Definition. \( \lim_{x \to c} \frac{1}{x} \) does not exist if for every number \( L \) there is at least one positive number \( \epsilon \) for which there is no positive number \( \delta \) such that \( |\frac{1}{x} - L| < \epsilon \) whenever \( |x| < \delta \).

Here is a proof:

Suppose we are given a number \( L \). Let \( \epsilon = L \). We would like to show that there is no positive number \( \delta \) such that \( |\frac{1}{x} - L| < \epsilon \) whenever \( |x| < \delta \), so suppose we are given an arbitrary \( \delta > 0 \). Consider \( x = \min(\delta/2, 1/(2|L|)) \). We know that \( 0 < x \leq \delta/2 \), so it is certainly true that \( |x| < \delta \).
However, we also have that \( x \leq 1/(2|L|) \), so \( 1/x \geq 2|L| \). Consequently, \( 1/x - L \geq |L| \) (because subtracting \( L \) cannot reduce the right-hand side by more than \( |L| \)). From this we can conclude that \( |1/x - L| \geq L \). Since there is at least one value of \( x \) for which \( |x| < \delta \) but not \( |1/x - L| < \epsilon \).
(remember that \( \epsilon = L \)) – and there is one such value for any possible value of \( \delta \) – we have proved that there is no positive number \( \delta \) such that \( |1/x - L| < \epsilon \) whenever \( |x| < \delta \). Since we showed that this is true for at least one value of \( \epsilon \) no matter what the value of \( L \) is, we have proved the statement in the definition.

The validity of the values of \( \epsilon \) and \( x \) produced in this proof is illustrated in figure 2.8.

Figure 2.8: Here, an example value of \( L = 1 \) was chosen. Our proof gives \( \epsilon = L = 1 \), which generates the blue bands centered around \( L \) in the graphs. Next an example value of \( \delta \) was chosen, which generates the red bands centered around \( c = 0 \) in the graphs. Our proof then gives \( x = \min(\delta/2, 1/(2|L|)) \), which is inside the red band without \( f(x) \) being inside the blue band – in both cases, just as we proved.

### 2.4.5 Pathological Function

Recall the function

\[
    f(x) = \begin{cases} 
        x & \text{if } x \text{ is rational} \\
        0 & \text{if } x \text{ is irrational}
    \end{cases}
\]

from page 5. It turns out that \( \lim_{x \to c} f(x) \) is 0 for \( c = 0 \) but the limit does not exist for \( c \neq 0 \). Why?

**Definition.** \( \lim_{x \to 0} f(x) = 0 \) if for every positive number \( \epsilon \) there is at least one positive number \( \delta \) such that \( |f(x)| < \epsilon \) whenever \( |x| < \delta \).

To prove this, simply let \( \delta = \epsilon \). If \( |x| < \delta \), then since \( |f(x)| \leq |x| \), we can use transitivity to conclude that \( |f(x)| < \epsilon \).

The fact that \( \lim_{x \to c} f(x) \) does not exist for \( c \neq 0 \) is a consequence of the fact that every possible interval of real numbers contains both rational numbers and irrational numbers. This means that any interval (of width \( \delta \), say) containing \( x = c \) will also contain some rational numbers greater than \( c \) and some irrational numbers greater than \( c \). Therefore, if we look at the definition of \( f(x) \), we
notice that in any interval containing \( x = c \), \( f(x) \) will be 0 at some points and greater than \( c \) at some points. All we have to do to prove that the limit does not exist, then, is to pick an \( \epsilon \) smaller than \( c/2 \), because no interval of height less than \( c \) will be able to include both 0 and values greater than \( c \).

### 2.5 Nuances

Our formal definition is currently as follows:

**Definition.** \( \lim_{x\to c} f(x) = L \) if for every positive number \( \epsilon \) there is at least one positive number \( \delta \) such that \( |f(x) - L| < \epsilon \) whenever \( |x - c| < \delta \).

Notice that it says \( |f(x) - L| < \epsilon \) should be true for *any* \( x \) where \( |x - c| < \delta \). But what about \( x = c \)? Our definition of a limit should not depend on \( f(c) \), because of functions like the second one shown in figure 2.1. (In fact, according to our current definition, only continuous functions can have limits!) So actually, we only want to consider values of \( x \) for which \( 0 < |x - c| < \delta \). The only reason we did not write this earlier was for conciseness. Therefore, a more correct definition is:

**Definition.** \( \lim_{x\to c} f(x) = L \) if for every positive number \( \epsilon \) there is at least one positive number \( \delta \) such that for all \( x \) in the domain of \( f \) that satisfy \( 0 < |x - c| < \delta \), the inequality \( |f(x) - L| < \epsilon \) holds.

The one final modification to our definition of limit involves the fact that \( \lim_{x\to 0} \sqrt{x} \) does not exist according to our current definition.

First, let’s discuss why this is the case. For this limit \( c = 0 \), so no matter what value is given to \( \delta \), there will be negative values of \( x \) that satisfy \( 0 < |x - c| < \delta \). (For instance, \( x = -\delta/2 \).) And if \( x \) is negative, then \( f(x) = \sqrt{x} \) is not defined, so the statement \( |f(x) - L| < \epsilon \) is not true. Therefore, no matter what values \( \epsilon \) and \( \delta \) have, it is never the case that \( |f(x) - L| < \epsilon \) whenever \( 0 < |x - c| < \delta \). And consequently, according to our definition, the limit does not exist.

Many mathematicians believe that the above limit *should* exist, so they modify the constraint on \( x \) to also require that \( x \) be within the domain of \( f \). This means we cannot pick a negative number for \( x \) to disprove the limit \( \lim_{x\to 0} \sqrt{x} = 0 \) like we did above. As you might have guessed from section 2.4.3, actually proving this limit would be quite unwieldy, but it is certainly possible.

If you (or your class) would like \( \lim_{x\to 0} \sqrt{x} \) to exist, then simply make the following change to the definition of limit:

**Definition.** \( \lim_{x\to c} f(x) = L \) if for every positive number \( \epsilon \) there is at least one positive number \( \delta \) such that for all \( x \) in the domain of \( f \) that satisfy \( |x - c| < \delta \), the inequality \( |f(x) - L| < \epsilon \) holds.

This is the definition we will use throughout the rest of the document, although the part about the domain won’t actually affect any of our later work.

### 2.6 Extending the Definition

Currently, we have a definition that tells us what it means for a function from real numbers to real numbers to limit to a real number. But since all of calculus is based on limits, when we want to do calculus in a different situation (such as describing the behavior of a function “at infinity” or analyzing a function of multiple variables) we must define a new type of limit for that situation.
2.6.1 Infinite Limits

You have most likely already seen statements such as

$$\lim_{x \to 0} \frac{1}{x^2} = \infty.$$ 

What does this statement mean? A normal limit statement such as

$$\lim_{x \to c} f(x) = L$$

means that $f(x)$ becomes arbitrarily close to $L$ as $x$ approaches $c$. An infinite limit statement such as

$$\lim_{x \to c} f(x) = \infty$$

means that $f(x)$ becomes arbitrarily large as $x$ approaches $c$.

We can specify a requirement for closeness with a single, small number $\epsilon$. Analogously, we can specify a requirement for largeness with a single, large number $N$. The requirement that $f(x)$ be within $\epsilon$ of $L$ is then satisfied when $|f(x) - L| < \epsilon$; analogously, the requirement that $f(x)$ be larger than $N$ is satisfied when $f(x) > N$. Therefore, we can change our original definition:

**Definition.** $\lim_{x \to c} f(x) = L$ if for every positive number $\epsilon$ there is at least one positive number $\delta$ such that for all $x$ in the domain of $f$ that satisfy $|x - c| < \delta$, the inequality $|f(x) - L| < \epsilon$ holds.

To this one:

**Definition.** $\lim_{x \to \infty} f(x) = \infty$ if for every number $N$ there is at least one positive number $\delta$ such that for all $x$ in the domain of $f$ that satisfy $|x - c| < \delta$, the inequality $f(x) > N$ holds.

2.6.2 Limits of Sequences

A sequence is a function $a_n$ whose domain is the positive integers $n$ (*not* real numbers). Often we are interested in seeing whether the values of $a_n$ approaches a limit $L$ as $n$ becomes very large. (This is written as $\lim_{n \to \infty} a_n = L$.) In this case, we have two substitutions to make. First, instead of a small number $\delta$ we will have a large number $M$. The requirement will then change from $|x - c| < \delta$ to $n > M$. As a result, the definition becomes:

**Definition.** $\lim_{x \to \infty} a_n = L$ if for every positive number $\epsilon$ there is at least one positive integer $M$ such that for all positive integers $n$ that satisfy $n > M$, the inequality $|f(x) - L| < \epsilon$ holds.

2.6.3 Limits in Multiple Dimensions

Sometimes we have a function of multiple variables $f(x, y)$ and would like to describe its limit as the point $(x, y)$ approaches a particular point $(x_0, y_0)$. This is written

$$\lim_{(x, y) \to (x_0, y_0)} f(x, y) = L.$$ 

In this case, we need to change the inequality $|x - c| < \delta$ because it only takes into account one variable. It is quite natural to extend it, however: $|x - c|$ is the distance between an arbitrary position $x$ and the limiting position $c$; analogously, $\sqrt{(x - x_0)^2 + (y - y_0)^2}$ is the distance between an arbitrary position $(x, y)$ and the limiting position $(x_0, y_0)$. Therefore, we can change the definition as follows:
Definition. \( \lim_{(x,y) \to (x_0,y_0)} f(x,y) = L \) if for every positive number \( \epsilon \) there is at least one positive number \( \delta \) such that for all \((x,y)\) in the domain of \( f \) that satisfy \( \sqrt{(x-x_0)^2 + (y-y_0)^2} < \delta \), the inequality \(|f(x) - L| < \epsilon \) holds.

### 2.6.4 Riemann Sums

Extending the limit definition to cover a mathematical object known as a *Riemann sum* is a key part of defining the integral, and is discussed in section 4.1.3.

### 3 Differentiation

#### 3.1 The Product Rule

**Prerequisite Knowledge.** Understand the definition of a derivative, and understand some simple limit rules.

In the product rule, one is interested in the derivative of the function \( h(x) = f(x)g(x) \). This is defined as the limit of the expression

\[
\frac{h(x + \Delta x) - h(x)}{\Delta x}
\]

as \( \Delta x \) approaches zero. Of course, according to the definition of \( h(x) \), we also have

\[
\frac{h(x + \Delta x) - h(x)}{\Delta x} = \frac{f(x + \Delta x)g(x + \Delta x) - f(x)g(x)}{\Delta x}.
\]

In order to simplify the notation and to make it more clear how the changes in different variables affect the overall expression, we will let \( f = f(x) \) and \( \Delta f = f(x + \Delta x) - f(x) \), so that \( f + \Delta f = f(x + \Delta x) \). (Note that as \( \Delta x \) approaches zero, so does \( \Delta f \).) Our derivative is then the limit of

\[
\frac{(f + \Delta f)(g + \Delta g) - fg}{\Delta x} = \frac{fg + f\Delta g + g\Delta f + \Delta f \Delta g - fg}{\Delta x} = \frac{\Delta g}{\Delta x} + \frac{\Delta f}{\Delta x} \Delta g.
\]

as \( \Delta x \) approaches zero. Since we are taking the limit of these expressions as \( \Delta x \) approaches zero, the expressions \( \Delta f/\Delta x \) and \( \Delta g/\Delta x \) become the derivatives \( f' \) and \( g' \) respectively. The derivative of \( fg \) is then

\[
fg' + gf' + f'\Delta g.
\]

Since we are assuming both of the derivatives exist, i.e. that they are just ordinary real numbers without any special infinite or discontinuous behavior, when we let \( \Delta x \) approach zero, so will \( \Delta g \) and so consequently will \( f'\Delta g \). Therefore, this term becomes zero when the limit is taken, giving

\[
(fg)' = fg' + gf'.
\]

#### 3.2 The Quotient Rule
**Prerequisite Knowledge.** Understand the definition of a derivative, and understand some simple limit rules.

In the quotient rule, one is interested in the derivative of the function \( h(x) = f(x)/g(x) \). This is defined as the limit of the expression

\[
\frac{h(x + \Delta x) - h(x)}{\Delta x}
\]

as \( \Delta x \) approaches zero. Of course, according to the definition of \( h(x) \), we also have

\[
\frac{h(x + \Delta x) - h(x)}{\Delta x} = \frac{\frac{f(x + \Delta x)}{g(x + \Delta x)} - \frac{f(x)}{g(x)}}{\Delta x}.
\]

In order to simplify the notation and to make it more clear how the changes in different variables affect the overall expression, we will let \( f = f(x) \) and \( \Delta f = f(x + \Delta x) - f(x) \), so that \( f + \Delta f = f(x + \Delta x) \). (Note that as \( \Delta x \) approaches zero, so does \( \Delta f \).) Our derivative is then the limit of

\[
\frac{f + \Delta f}{g + \Delta g} - \frac{f}{g} = \frac{\frac{f(x + \Delta x)g(x) - f(x)g(x + \Delta x)}{g(x + \Delta x)g(x)}}{\Delta x} = \frac{g\Delta f - f\Delta g}{(g^2 + g\Delta g)\Delta x} = \frac{g\Delta f}{g^2 + g\Delta g}.
\]

Since we are taking the limit of this expression as \( \Delta x \) approaches zero, the expressions \( \Delta f/\Delta x \) and \( \Delta g/\Delta x \) become the derivatives \( f' \) and \( g' \) respectively. The derivative of \( f/g \) is then

\[
\frac{gf' - fg'}{g^2 + g\Delta g}.
\]

Also, when we let \( \Delta x \) approach zero, so will \( \Delta g \) and so consequently will \( g\Delta g \). Therefore, the term \( g^2 + g\Delta g \) becomes simply \( g^2 \), giving

\[
\left( \frac{f}{g} \right)' = \frac{gf' - fg'}{g^2}.
\]

4 **Integration**

4.1 **The Formal Definition of the Integral**

**Prerequisite Knowledge.** Understand the formal definition of limit (section 2) and in particular how it is extended (section 2.6).

4.1.1 **The Idea**

You should know that

\[
\int_a^b f(x) \, dx
\]

corresponds to the area bounded to the left by \( x = a \), to the right by \( x = b \), below by \( y = 0 \), and above by \( y = f(x) \) – assuming that \( f \) is both positive and continuous. However, just as with the formal definition of limit (section 2), although it is very valuable to have a good intuitive understanding (to quickly solve simple problems), it is also valuable to have a good rigorous understanding (to accurately solve complex problems).
To this end, we will invent a formal definition for the integral which gives us the area under a curve for a positive, continuous function. Then we can apply the same definition to other, less well-behaved functions and we can obtain possibly useful information about those functions.

By the way, there is more than one way to define the integral (in particular, a popular way uses upper and lower Riemann sums), but I think that this way is the easiest to understand. It should be easy to learn another definition once you understand how to understand this type of definition.

4.1.2 Calculating Area

![Figure 4.1: The area under the graph of $y = f(x)$ between $x = a$ and $x = b$.](image)

It is unlikely that you learned a formula in geometry for finding the shaded area under the graph of $y = f(x)$ in figure 4.1, especially given that you don’t know what $f(x)$ is. On the contrary, it is extremely likely you learned a formula for finding the area of a rectangle, and we can approximate any shape with a set of rectangles.

For example, we can approximate the shaded area in figure 4.1 with a single rectangle, which is shown in figure 4.2. It should seem reasonable that the rectangle we use has a width of $b - a$, but what about the height? Theoretically, depending on the function, many different heights might be good estimates. It’s clearly absurd, however, to make the rectangle taller or shorter than the function ever reaches. Therefore, we will pick a number $c$ between $a$ and $b$, and the height of the rectangle will be $f(c)$. The area given by this (bad) approximation is $A = f(c)[b - a]$.

It would most assuredly be better to use two rectangles in our approximation. We can do this by splitting the interval $[a, b]$ into two subintervals, then approximating each subinterval with a single rectangle. This is shown in figure 4.3. In the figure, we have introduced some more consistent notation for the various $x$-coordinates in order to make things easier to keep track of. It should be fairly easy to see that the total shaded green area in the figure is $A = f(c_1)[x_1 - x_0] + f(c_2)[x_2 - x_1]$. Note, by the way, that there are three choices to make in approximating area with two rectangles: we must choose a number $x_1$ between $a$ and $b$, we must choose a number $c_1$ between $x_0$ and $x_1$, and we must choose a number $c_2$ between $x_1$ and $x_2$. To write the formula for area more concisely, we can define $\Delta x_1 = x_1 - x_0$ and $\Delta x_2 = x_2 - x_1$. This gives us $A = f(c_1)\Delta x_1 + f(c_2)\Delta x_2$. 

20
Figure 4.2: An approximation of the area under the graph of $y = f(x)$ using one rectangle.

$\sum_{i=1}^{N-1} f(c_i)(x_{i+1} - x_i)$

Figure 4.3: An approximation of the area under the graph of $y = f(x)$ using two rectangles.

In general, we can make an approximation of the shaded area in figure 4.1 using $N$ rectangles. This will involve choosing $N - 1$ points to define the borders of the rectangles: the first rectangle is from $a = x_0$ to $x_1$, the second is from $x_1$ to $x_2$, and so on until the last rectangle is from $x_{N-1}$ to $x_N = b$. Then we will have to choose $N$ points, one within each rectangle, to define the heights of the rectangles: the first rectangle will have a height of $f(c_1)$, where $c_1$ is between $x_0$ and $x_1$, and so on. This setup, for $N = 5$ rectangles, is shown in figure 4.4. The area of the green region in the figure is then

$$A = f(c_1)[x_1 - x_0] + f(c_2)[x_2 - x_1] + f(c_3)[x_3 - x_2] + f(c_4)[x_4 - x_3] + f(c_5)[x_5 - x_4]$$

$$= f(c_1) \Delta x_1 + f(c_2) \Delta x_2 + f(c_3) \Delta x_3 + f(c_4) \Delta x_4 + f(c_5) \Delta x_5$$
\[ S = \sum_{i=1}^{5} f(c_i) \Delta x_i. \]

Figure 4.4: An approximation of the area under the graph of \( y = f(x) \) using \( N = 5 \) rectangles.

### 4.1.3 Using Limits

So now we have a formula for approximating the area under the graph of a function,

\[ S = \sum_{i=1}^{N} f(c_i) \Delta x_i. \]

This formula is called a **Riemann sum**, and its value depends on the number of subintervals \( (N) \), the points at which we partition the interval \( (x_1, x_2, \ldots, x_{N-1}) \), and the points we pick in each subinterval to determine the heights of the rectangles \( (c_1, c_2, \ldots, c_N) \). Now, none of these approximations actually give the correct answer\(^2\), but you can see that the more rectangles we use the closer our approximation will get to the true area. So what we really want is a way of describing the value to which our approximations get closer and closer. This idea is almost identical to that of a limit, as was discussed in section 2. The definition is reprinted here for convenience:

**Definition.** \( \lim_{x \to c} f(x) = L \) if for every positive number \( \epsilon \) there is at least one positive number \( \delta \) such that for all \( x \) in the domain of \( f \) that satisfy \( |x - c| < \delta \), the inequality \( |f(x) - L| < \epsilon \) holds.

It is not too difficult to adapt this definition to describe the limit of a Riemann sum rather than a function, analogously to our work in section 2.6. Firstly, \( f(x) \) must be replaced with \( S \). However, \( |x - c| < \delta \) no longer makes any sense: \( f(x) \) depended only on \( x \), but \( S \) depends on a great many different numbers.

\(^2\)Actually, a select few of the approximations will give the correct answer, according to the mean value theorem for integrals. But we can’t use that theorem until we define what an integral means!
The question in the original definition was: what condition forces \( f(x) \) to approach \( L \)? The answer was: \( x \) must become arbitrarily close to \( c \), or \( |x - c| < \delta \). The new question is: what condition forces \( S \) to approach \( L \)? In other words, what condition forces our rectangular approximations to become arbitrarily accurate? Your first instinct might be to say that the number of rectangles must become arbitrarily large, or \( N > M \), but there is a subtlety that prevents this from working.

![Figure 4.5: A (bad) rectangular approximation of the area under \( y = \sqrt{x} \) from \( x = 0 \) to \( x = 4 \) with the property that it never becomes a good approximation no matter how many rectangles are used.](image)

In figure 4.5, the borders of the rectangles are at the \( x \)-coordinates: 4, 2, 1, \( 1/2 \), \( 1/4 \), \( 1/8 \), \ldots, \( 1/2^n \). Each evaluation point \( c_i \) is the right-most point within its rectangle. You can easily see that the first rectangle substantially overestimates the area under the curve \( y = \sqrt{x} \), and when we add new rectangles they are just added to the left-hand side (they do not affect the first rectangle). Therefore, that overestimation will never go away, and so \( S \) will never approach the true area under the curve \( (L) \).

So we need a different way to specify that the rectangular approximation becomes arbitrarily accurate. It turns out that the best way to do this is to specify that the width of the largest rectangle becomes arbitrarily small. This will exclude the approximation in figure 4.5 because the width of the largest rectangle never drops below 2. If we denote the width of the largest rectangle (or, equivalently, the greatest \( \Delta x_i \)) as \( ||\Delta|| \), then we can replace \( |x - c| < \delta \) (\( x \) becomes arbitrarily close to \( c \)) with \( ||\Delta|| < \delta \) (\( ||\Delta|| \) becomes arbitrarily small).

With a few other things cleaned up, we get the following definition for the limit of a Riemann sum, which I like to write as

\[
\lim_{||\Delta|| \to 0} \sum_{i}^{a \to b} f(c_i)\Delta x_i = L.
\]

(The symbol \( \sum_{i}^{a \to b} \) is just shorthand to indicate a summation \( \sum_{i=1}^{N} \) where the first \( x_i \) is \( a \) and the last \( x_i \) is \( b \).)

**Definition.** \( \lim_{||\Delta|| \to 0} \sum_{i}^{a \to b} f(c_i)\Delta x_i = L \) if for every positive number \( \epsilon \) there is at least one positive number \( \delta \) such that for all Riemann sums of \( f \) over \([a, b]\) that satisfy \( ||\Delta|| < \delta \), the inequality \( |S - L| < \epsilon \) holds.

And there we have our definition: if

\[
\lim_{||\Delta|| \to 0} \sum_{i}^{a \to b} f(c_i)\Delta x_i = L,
\]

then

\[
\int_{a}^{b} f(x) \, dx = L.
\]
4.1.4 Implications

Unfortunately, this definition of the integral does not give us a simple way to actually find its value, but it rather gives us an extraordinarily difficult limit to try to guess and then subsequently prove formally. Luckily, there are two facts that make it possible to sidestep the limit definition somewhat:

- For continuous, positive functions, the integral corresponds to the area under the curve. This is simply a consequence of the fact that we derived our definition from finding the area under a function’s graph. For example, \( \int_{-1}^{1} \sqrt{1-x^2} \, dx = \pi/2 \) because the area under the graph is a semicircle of radius 1.

- There is a theorem stating that if \( f \) is continuous on \([a, b]\), then the limit in our definition of \( \int_{a}^{b} f(x) \, dx \) exists. This should make intuitive sense – all it is saying is that any smooth (continuous) curve can be approximated as well as you might want using rectangles.

The second fact may not seem very important, but it is the key to evaluating integrals. Why? If your function is continuous, then the theorem tells you that the integral for its integral exists (which is very hard to prove in general) – all you have to do is find the actual value (which is usually much easier).

To see how to use the limit definition to evaluate an integral, consider \( \int_{0}^{1} x \, dx \).

Now consider the Riemann sum consisting of \( N \) rectangles of equal widths (since the total interval has width 1, each rectangle has a width of \( 1/N \)) and with each point of evaluation being at the far right edge of its rectangle. Then rectangle \( i \) is bounded on the left by \( x = (i-1)(1/N) \) and bounded on the right by \( i/N \). This means that \( x_i = c_i = i/N \), and \( \Delta x_i = x_i - x_{i-1} = i/N - (i-1)/N = 1/N \).

We can then directly find the value of this Riemann sum:

\[
S_n = \sum_{i=1}^{N} f(c_i) \Delta x_i = \sum_{i=1}^{N} f \left( \frac{i}{N} \right) \left[ \frac{1}{N} \right] = \sum_{i=1}^{N} \frac{i}{N^2} = \frac{1}{N^2} \sum_{i=1}^{N} i = \frac{1}{N^2} \left( \frac{N(N + 1)}{2} \right)
\]

\[
= \frac{N + 1}{2N} = \frac{1}{2} + \frac{1}{2N}
\]

Now, since \( f(x) = x \) is continuous, we know that \( \lim_{||\Delta|| \to 0} \sum_{i}^{a \to b} f(c_i) \Delta x_i \) exists. What does this mean? According to the definition, it means that \( S \) gets arbitrarily close to \( L \) as \( ||\Delta|| \) gets arbitrarily close to 0. But wait: we already have a sequence of Riemann sums for which \( ||\Delta|| \) gets arbitrarily close to 0: the sums we computed above, for which \( ||\Delta|| = \Delta x_i = 1/N \) since all of the rectangles are of the same width. Clearly, as \( N \) approaches \( \infty \), \( ||\Delta|| = 1/N \) will approach 0. So \( S_n \) must get arbitrarily close to \( L \) as \( N \) approaches \( \infty \); in other words,

\[
\lim_{N \to \infty} S_n = L = \lim_{||\Delta|| \to 0} \sum_{i}^{a \to b} f(c_i) \Delta x_i = \int_{0}^{1} x \, dx.
\]

We have reduced the extremely complicated limit on the right to the comparatively trivial limit on the left, which is easily evaluated:

\[
\lim_{N \to \infty} S_n = \lim_{N \to \infty} \left[ \frac{1}{2} + \frac{1}{2N} \right] = \frac{1}{2}.
\]
Consequently,

$$\int_0^1 x \, dx = \frac{1}{2}.$$  

Of course, you could also evaluate this integral by noting that it represents a triangle with both base and height of 1. Geometrically, the quantity 1/2 + 1/2N represents the area of the triangle plus the overestimation accrued by using rectangles, which decreases as the number of rectangles N increases.

But you probably use the fundamental theorem of calculus to evaluate integrals, so why bother with all these details about Riemann sums? Well, in fact, the formal definition of the integral shows up in the proof of the fundamental theorem of calculus! See section 4.2.3 for details on this.

### 4.2 The First Fundamental Theorem of Calculus

**Prerequisite Knowledge.** Know what the symbols in the fundamental theorem of calculus mean, and understand the mean value theorem.

#### 4.2.1 The Form of the Equation

The fundamental theorem of calculus states that when the derivative of a function is integrated, the result is the original function evaluated at the bounds of integration. In other words,

$$\int_a^b f'(x) \, dx = f(b) - f(a).$$

Of course, if the function $f(x)$ is $G(x)$, an antiderivative of $g(x)$, then $f'(x) = G'(x) = g(x)$ and we get the more familiar result that

$$\int_a^b g(x) \, dx = G(b) - G(a).$$

However, when trying to understand the theorem intuitively, the first form is more useful. (It is difficult to visualize an antiderivative.)

#### 4.2.2 The Main Idea

$$\int_a^b \text{slope} \, dx = \text{change in } y$$

When slope (change in $y$ per change in $x$) is multiplied by $dx$, a small change in $x$, we get $dy$, a small change in $y$. Since $\int_a^b$ means “add a lot of, from $a$ to $b$”, when it is applied to $dy$ we get $\Delta y$, a large change in $y$ (from $a$ to $b$).

This is the fundamental idea of the theorem, but there is also a third component to keep in mind, which is that the left side of this equation also corresponds to the area under the graph of “slope” (also known as $f'$). This is because “slope”, at any given point, is the distance straight up from the $x$-axis to the graph of “slope”, and $dx$ is a small width. When we multiply these two numbers together, we get the area of a very thin rectangle extending from the $x$-axis to the graph of “slope”. This rectangle is a great approximation of the area under the graph in the small interval represented by $dx$. When we apply $\int_a^b$ (add a lot of, from $a$ to $b$) to the areas under the graph over small intervals, we get the total area under the graph in the whole interval (from $a$ to $b$).
4.2.3 The Details

Consider the graph of the function \( y = f(x) \) in figure 4.6.

![Graph of function](image)

Figure 4.6: A partition of the interval from \( x = a \) to \( x = b \) into three subintervals, with the values of the function \( y = f(x) \) shown at each boundary between subintervals. The slopes of the secant lines in each subinterval are designated \( m_1, m_2, \) and \( m_3. \)

The most important quantity shown in the figure is \( \Delta y \), the difference between the \( y \)-coordinates of the first and last points. These points are located at \((a, f(a))\) and \((b, f(b))\) respectively, so \( \Delta y = f(b) - f(a) \). Also, the figure shows that \( \Delta y = \Delta y_1 + \Delta y_2 + \Delta y_3 \). In other words, the total change in \( y \) over all three subintervals is equal to the sum of the changes in \( y \) over each subinterval.

Now consider the secant lines in each subinterval, which have the slopes \( m_1, m_2, \) and \( m_3 \) respectively. For the first subinterval, the changes in \( x \) and \( y \) between the first and second points are shown on the figure as \( \Delta x_1 \) and \( \Delta y_1 \). For the secant line in the first subinterval, we know that \( m_1 = \Delta y_1 / \Delta x_1 \), so \( \Delta y_1 = m_1 \Delta x_1 \). After doing the same for the second and third subintervals, we can rewrite the equation for \( \Delta y \) as \( \Delta y = m_1 \Delta x_1 + m_2 \Delta x_2 + m_3 \Delta x_3 \). Already, the fundamental idea is clear: performing a summation over slope multiplied by small change in \( x \) gives total change in \( y \).

To be formal, however, we can proceed with the mean value theorem. [We are assuming that the function is differentiable – otherwise, how could we even have an \( f'(x) \) to use in the formula of the fundamental theorem of calculus?] For the first subinterval, it tells us that for some \( x \) (call it \( c_1 \)) between \( x_1 \) and \( x_2 \), the slope of the tangent line at \( c_1 \) is equal to the slope of the secant line from \((x_1, y_1)\) to \((x_2, y_2)\). In other words, there is a \( c_1 \) in the first subinterval such that \( f'(c_1) = m_1 \).
Since the mean value theorem applies just as well to each other subinterval, we also have a $c_2$ in the second subinterval and a $c_3$ in the third subinterval such that $f'(c_2) = m_2$ and $f'(c_3) = m_3$. We can then rewrite the equation for $\Delta y$ again as $\Delta y = f'(c_1)\Delta x_1 + f'(c_2)\Delta x_2 + f'(c_3)\Delta x_3$. Since this equation has quite a bit of repetition, we can also write it as a summation:

$$\Delta y = \sum_{i=1}^{3} f'(c_i)\Delta x_i.$$  

Recall from earlier that $\Delta y = f(b) - f(a)$. Also, the equation we found is true for any number of subintervals, not just 3. If we let $N$ be the number of subintervals, then we get

$$f(b) - f(a) = \sum_{i=1}^{N} f'(c_i)\Delta x_i.$$  

Really, this equation is the fundamental theorem of calculus, except that it is for a finite number of subintervals instead of an infinite number. If we take the limit of both sides as the number of subintervals approaches infinity,

$$\lim_{N \to \infty} [f(b) - f(a)] = \lim_{N \to \infty} \sum_{i=1}^{N} f'(c_i)\Delta x_i,$$

then the left side stays the same (it does not depend upon the number of subintervals), while the right side is exactly the definition of an integral.\(^3\) Therefore, we have

$$f(b) - f(a) = \int_{a}^{b} f'(x) \, dx.$$  

### 4.2.4 Alternate Proof

**Prerequisite Knowledge.** Understand Euler’s method.

We can also consider the fundamental theorem of calculus as “Euler’s method with an infinitely small step size”. As a visual aid, figure 4.7 shows the approximation of a function using Euler’s method. As in figure 4.6, the graph shows $f(x)$, while we examine $f'(x)$ in the equations.

Just as in figure 4.6, the first and last $x_i$ are the bounds; i.e., $x_1 = a$ and $x_5 = b$. The point of Euler’s method is to approximate the change $\Delta y = f(b) - f(a)$ using only information about the derivative $f'(x)$. This approximation is:

$$\Delta y \approx y_5 - y_1 = (y_2 - y_1) + (y_3 - y_2) + (y_4 - y_3) + (y_5 - y_4) = f'(x_1)[x_2 - x_1] + f'(x_2)[x_3 - x_2] + f'(x_3)[x_4 - x_3] + f'(x_4)[x_5 - x_4] = f'(x_1)\Delta x_1 + f'(x_2)\Delta x_2 + f'(x_3)\Delta x_3 + f'(x_4)\Delta x_4 = \sum_{i=1}^{4} f'(x_i)\Delta x_i.$$  

---

\(^3\)Since $f(x)$ is differentiable, $f'(x)$ is continuous (in all but the most unusual of cases). Therefore, according to our discussion in section 4.1.4, the integral exists and all we have to do is find one sequence of Riemann sums such that $\|\Delta\| \to 0$ in order to find the value of the integral.

\(^4\)Consider the function defined so that $f(x) = x^2 \sin(1/x)$ for $x \neq 0$ and so that $f(0) = 0$. 

27
Figure 4.7: Euler’s method, using the same naming conventions as figure 4.6. The graph of \( y = f(x) \) is shown, and Euler’s method is applied to the derivative \( f'(x) \) in order to find the slope of each line segment.

Now, Euler’s method would be no good if its approximation did not get better and better as we used more and more steps. So, as a reminder, we have

\[
f(b) - f(a) \approx \sum_{i=1}^{N} f'(x_i) \Delta x_i
\]

for any finite number of steps \( N \). However, if we take the limit as \( N \to \infty \), then the approximation will become correct and we find once again that

\[
f(b) - f(a) = \lim_{N \to \infty} \sum_{i=1}^{N} f'(x_i) \Delta x_i = \int_{a}^{b} f'(x) \, dx.
\]

Note that once again, the key idea is multiplying slope \( f'(x) \) by change in \( x \) to get change in \( y \), and then summing all of those small changes in \( y \) to get the total \( f(b) - f(a) \).

Really, numerical approximation of an integral using a left Riemann sum and numerical approximation of a differential equation using Euler’s method are exactly the same thing (just with different names and different intended uses). This can also be seen by the fact that the way to solve the differential equation

\[
\frac{dy}{dx} = f'(x)
\]

is to integrate both sides with respect to \( x \) – thus intimately relating the solutions of integrals and first-order differential equations.

4.2.5 Conclusion

\[
\int_{a}^{b} f'(x) \, dx = \int_{a}^{b} \frac{dy}{dx} \, dx = \int_{f(a)}^{f(b)} dy = f(b) - f(a).
\]
### 4.3 The Second Fundamental Theorem of Calculus

**Prerequisite Knowledge.** Know what the symbols in the second fundamental theorem of calculus mean.

The second fundamental theorem of calculus is often presented as

\[
\frac{d}{dx} \int_a^x f(t) \, dt = f(x).
\]

However, the important idea is that you are finding the rate of change of the value of the integral with respect to its upper bound. We can relabel some variables to make this more clear:

\[
\frac{d}{db} \int_a^b f(x) \, dx = f(b).
\]

Now consider the meaning of the various symbols in the formula. Firstly, \(\int_a^b f(x) \, dx\) is the area under the graph of \(y = f(x)\) from \(x = a\) to \(x = b\). \(d/db\) signifies rate of change with respect to \(b\). In this case, the left side of the formula is the rate of change of the area with respect to its right-hand bound. Mathematically, this means

\[
\frac{\Delta \text{[area]}}{\Delta b} = \frac{[\text{area from } a \text{ to } (b + \Delta b)] - [\text{area from } a \text{ to } b]}{\Delta b} = \frac{[\text{area from } b \text{ to } (b + \Delta b)]}{\Delta b}.
\]

If \(\Delta b\) is small, then the value of the function will not change significantly over the interval \([b, b + \Delta b]\), but will remain approximately \(f(b)\). The area under the curve is then approximately a rectangle with height \(f(b)\) and width \(\Delta b\), with area \(f(b)\Delta b\). We then have

\[
\frac{d}{db} \int_a^b f(x) \, dx = \frac{\Delta \text{[area]}}{\Delta b} = \frac{f(b)\Delta b}{\Delta b} = f(b).
\]

Another (less intuitive) way of proving the second fundamental theorem of calculus is to use the first fundamental theorem of calculus,

\[
\int_a^b f(x) \, dx = F(b) - F(a).
\]

If we differentiate both sides of this equation with respect to \(b\), \(F(b)\) will become \(F'(b) = f(b)\), while \(F(a)\) will be a constant. Therefore, we have

\[
\frac{d}{db} \int_a^b f(x) \, dx = \frac{d}{db} \left[ F(b) - F(a) \right] = f(b).
\]

### 5 Sequences and Series

Formal proofs can be given for each of the theorems in this section. They are not given here, and instead each theorem is accompanied by an explanation of why the result makes sense. This, I believe, is far more helpful, especially in remembering all of the conditions that are associated with the various theorems. (If you know, intuitively, why a theorem is true, then it is always easy to recall what its conditions are.)

Also, in this section I often refer to the “vergence” of a series or integral. This means the fact of its converging or diverging, as in: the two series have the same vergence; this diagram helps you understand the vergence of the series; we can determine the integral’s vergence.
5.1 Bounded Monotonic Sequences

**Prerequisite Knowledge.** Know what a sequence is and what notation is used for sequences.

This section discusses sequences, not series.

Suppose we have a sequence $a_n$, and that we know two things about it: (1) its terms never decrease, and (2) its terms never become greater than a certain number (call it $M$, for maximum). The graph of such a sequence is depicted in figure 5.1.

![Figure 5.1](image.png)

Figure 5.1: A nondecreasing sequence that is bounded above. No term in the sequence is greater than 4, including the ones not shown on the graph. This sequence happens to be $a_n = 4 - 3\exp(-n/4)$.

It is reasonable to conclude that this sequence would have to converge, because it cannot diverge in any of the usual ways sequences diverge. It cannot diverge by decreasing without bound (because it is nondecreasing), it cannot diverge by increasing without bound (because it is bounded above), and it cannot diverge by oscillating (because it is nondecreasing and oscillations require both increase and decrease).

To be precise, the number $M$ which is greater than all terms in the sequence is called an *upper bound*. Of course, there are (infinitely) many upper bounds for a sequence that is bounded above. For the sequence in figure 5.1, for instance, both 4 and 5 are upper bounds. However, 4 is more interesting because it is the *least upper bound*: any number less than 4 will no longer be an upper bound. In fact, according to a property of the real numbers called *completeness*, any sequence that has an upper bound also has a least upper bound.

You can use this fact to conclude that a nondecreasing sequence bounded above converges. Firstly, you know that the sequence will get arbitrarily close to its least upper bound (call it $L$, for least). If it did not, then there would be some difference between $L$ and the highest the sequence reaches, which would mean that an upper bound smaller than $L$ would exist. This would contradict the fact that $L$ is the least upper bound. You also know that the sequence can only get closer to $L$, not farther away, because it is nondecreasing. Since the sequence gets arbitrarily close to $L$ and does not ever get farther away from $L$, clearly its limit must be $L$. (What a coincidental choice of variable name!) Since the sequence has a limit, it converges.

We saw that a nondecreasing sequence bounded above must converge, and, analogously, a non-increasing sequence bounded below must also converge. Therefore, we know that:
**Theorem 1.** If a sequence is nondecreasing and bounded above, or it is nonincreasing and bounded below, then it converges.

Since *bounded* means “both bounded above and bounded below”, and since *monotonic* means “either nonincreasing or nondecreasing”, the above theorem also implies that all bounded monotonic sequences converge, which is the result typically presented in textbooks.

### 5.2 The Integral Test and Remainder Formula

**Prerequisite Knowledge.** Understand what improper integrals and infinite series look like graphically.

![Figure 5.2: The graph of a function $y = f(n)$ to which the integral test may be applied. The area under the curve is overestimated by the light gray rectangles (a left-hand approximation) and underestimated by the dark gray rectangles (a right-hand approximation). Only eight rectangles are shown, but there are (infinitely) more to the right.](image)

An infinite series and an improper integral are really quite similar. Both relate to areas on the graph of a function, as shown in figure 5.2. In the figure, the area of the light gray rectangle with a left side at $x = n$ is $f(n)$, and the area of the dark gray rectangle with a left side at $x = n$ is $f(n + 1)$. Therefore, the sum of the areas of the light gray rectangles is

$$
\sum_{n=1}^{\infty} f(n),
$$

and the sum of the dark gray rectangles is

$$
\sum_{n=1}^{\infty} f(n + 1) = \sum_{n=2}^{\infty} f(n).
$$

Since the area under the curve (right of $x = 1$) is less than the area of the light gray rectangles but greater than the area of the dark gray rectangles, we can write the inequality

$$
\sum_{n=2}^{\infty} f(n) \leq \int_1^{\infty} f(n) \, dn \leq \sum_{n=1}^{\infty} f(n).
$$
This is simply a statement that left-hand Riemann approximations overestimate integrals and right-hand Riemann approximations underestimate them, for decreasing functions. If the function in the graph were not decreasing, then it could easily have peaks or valleys that could bring it below the dark gray rectangles or above the light gray rectangles, as shown in figure 5.3.

Figure 5.3: A function to which the integral test does not apply. Its graph illustrates that the rectangles used in the derivation of the integral test are not necessarily circumscribed and inscribed (as they need to be) unless the function is decreasing, or at least nonincreasing.

In essence, these inequalities tell us that the integral \( \int_1^\infty f(n) \, dn \) and the summation \( \sum_{n=1}^{\infty} f(n) \) do not differ from each other by more than the finite term \( f(1) \). Since they are so close together, they must either both be finite or both be infinite (both converge or both diverge). This is the integral test:

\[ \text{Theorem 2.} \quad \text{If a function } f(n) \text{ is nonincreasing and integrable, then } \int_1^\infty f(n) \, dn \text{ and } \sum_{n=1}^{\infty} f(n) \text{ either both converge or both diverge.} \]

To reason rigorously: If the series on the right diverges, then the series on the left will also diverge since it only differs by a single, finite term. Since the integral is greater than the series on the left, it will also diverge. Similarly, if the series on the right converges, the integral must also converge since it is smaller. This shows that the vergence of the integral and the summation must be the same.

There are two other conditions that are often given as requirements for the integral test. In fact, neither of these conditions are necessary.

- The function must be continuous: It is only necessary for the function to be integrable (because its integral is used in the theorem), and the proof we followed will work just as well for non-continuous functions (as long as they are nonincreasing, of course).

- The function must be positive: Consider what would happen if the function were not always positive. Since it is nonincreasing, once the function becomes negative, it can never approach zero because that would involve increasing. We know that all series whose terms do not approach zero diverge, and similarly, if the magnitude of a function never approaches zero, its improper integral will diverge. This is the conclusion of the integral test (that the series and the integral have the same vergence). So even if the function is not always positive, the integral test still gives the correct result (as long as the function is nonincreasing). However, if a function is nonincreasing and not always positive, it will always diverge (by the \( n \)th term
test). It is therefore rather pointless to apply the integral test. (However, there’s still no point in adding one more condition to remember if it’s not even necessary: that just adds confusion!)

Since improper integrals are typically easier to evaluate than infinite series, the inequality of the integral test (reprinted here for convenience)

\[
\sum_{n=2}^{\infty} f(n) \leq \int_{1}^{\infty} f(n) \, dn \leq \sum_{n=1}^{\infty} f(n)
\]

also gives us a nice approximation for nonincreasing infinite series. From the two right-hand terms, we have

\[
\int_{1}^{\infty} f(n) \, dn \leq \sum_{n=1}^{\infty} f(n),
\]

and we can add \( f(1) \) to each of the two left-hand terms to get

\[
\sum_{n=1}^{\infty} f(n) \leq f(1) + \int_{1}^{\infty} f(n) \, dn.
\]

Combining these two inequalities, we have that

\[
\int_{1}^{\infty} f(n) \, dn \leq \sum_{n=1}^{\infty} f(n) \leq f(1) + \int_{1}^{\infty} f(n) \, dn.
\]

However, this only bounds the summation within an interval of width \( f(1) \). To get a better approximation, we can re-write the integral test inequality, except placing the first rectangle at \( n = N \) instead of \( n = 1 \):

\[
\sum_{n=N+1}^{\infty} f(n) \leq \int_{N}^{\infty} f(n) \, dn \leq \sum_{n=N}^{\infty} f(n).
\]

Now, following the same steps as before except adding \( f(N) \) instead of \( f(1) \) to get the inequality on the left-hand side, we have

\[
\int_{N}^{\infty} f(n) \, dn \leq \sum_{n=N}^{\infty} f(n) \leq f(N) + \int_{N}^{\infty} f(n) \, dn.
\]

This bounds part of the series within an interval of width \( f(N) \). Of course, since the function is decreasing, this is a smaller interval than before, which makes a better approximation. To get a bound for the entire summation, just add \( \sum_{n=1}^{N-1} f(n) \) to all three terms. We then find the integral test error bound:

**Theorem 3.** For any nonincreasing, integrable function \( f \) and positive integer \( N \),

\[
\sum_{n=1}^{N} f(n) + \int_{N}^{\infty} f(n) \, dn \leq \sum_{n=1}^{\infty} f(n) \leq \sum_{n=1}^{N-1} f(n) + \int_{N}^{\infty} f(n) \, dn.
\]

This inequality allows you to add a finite number of terms of a series and compute an integral, then use these numbers to place a bound on the value of a nonincreasing series (of which presumably you cannot find the exact value).
5.3 The Limit Comparison Test

**Prerequisite Knowledge.** Understand limits and series intuitively, and be able to evaluate limits of simple quotients.

Suppose we have two series, \( a_n \) and \( b_n \), such that

\[
\lim_{n \to \infty} \frac{a_n}{b_n} = C,
\]

where \( C \) is a nonzero but finite number. For now, suppose that every term of both series is positive. (We will consider otherwise later.) What this limit says is that

\[
\frac{a_n}{b_n} \approx C \text{ (for sufficiently large } n \text{)},
\]

or equivalently

\[
a_n \approx b_n \cdot C \text{ (for sufficiently large } n \text{)}.
\]

Now, if we have a convergent series and multiply it by a finite number, it will remain convergent; similarly, if we have a divergent series and multiply it by a finite number, it will remain divergent – as long as the number is nonzero, of course! So it follows directly that if \( a_n/b_n = C \) for all \( n \), then the series must have the same convergence. The limit comparison test is simply an extension for the case that this equation is not *exactly* true, but is *approximately* true for large \( n \). We can disregard small \( n \), because any finite number of terms of a series will only sum to a finite number, and this cannot affect the convergence of a series. This gives us:

**Theorem 4.** If \( \lim_{n \to \infty} a_n/b_n = C \), where is \( C \) is finite and nonzero, and \( a_n \) and \( b_n \) are both positive for all (or at least all sufficiently large) \( n \), then \( a_n \) and \( b_n \) either both converge or both diverge.

Naturally, the next question is: why must all of the terms of both series be positive? The issue is a subtlety with conditionally convergent series. More specifically, we can form two series as follows:

\[
a_n = [\text{large but convergent series}] + [\text{small but divergent series}]
\]

and

\[
b_n = [\text{large but convergent series}].
\]

Clearly, \( a_n \) is divergent and \( b_n \) is convergent due to the way series add, so it would be a big problem if the limit comparison test were to apply to this pair of series. (That is, it would be a problem if \( \lim_{n \to \infty} a_n/b_n \) were finite and nonzero.) Now, your first thought might be for the large series to be, for instance, \( 1000/n^2 \) and for the small series to be \( 1/n \). However, the limit is

\[
\lim_{n \to \infty} \frac{1000/n^2 + 1/n}{1000/n^2} = \infty
\]

because \( 1/n \) eventually becomes larger than \( 1000/n^2 \). (You should verify both parts of this statement yourself.) This is not unexpected – after all, \( 1/n \) is divergent, while \( 1000/n^2 \) is convergent. In fact, this would always happen if it were not for alternating series. Recall that the only condition for convergence of an alternating series is that its terms decrease to zero (see section 5.4), which is far less stringent than the condition for convergence of a general series. For instance, \( (-1)^n/n^{3/2} \) is
convergent, and is also larger than the divergent $1/n$ in the limit. (Check this.) The key is that negative terms allow convergent series to be larger than divergent series in the limit. For this pair of series, we have

$$\lim_{n \to \infty} \frac{(-1)^n/\sqrt{n} + 1/n}{(-1)^n/\sqrt{n}} = 1.$$  

(You should also verify this limit.) This would contradict the conclusion of the limit comparison test because the series in the numerator and denominator have different vergence. So, due to the fact that conditionally convergent series can outweigh divergent series in the limit, we must either discard the idea of testing vergence by comparing the magnitudes of terms or discard series with negative terms from the limit comparison test. (Most people choose the latter.)

5.4 The Alternating Series Test and Remainder Formula

**Prerequisite Knowledge.** Understand series and their notation.

An example of an alternating series is $\sum_{n=1}^{\infty} (-1)^{n+1}a_n$, where $a_n$ is always positive. The best way to understand the vergence of this series is to draw a diagram such as figure 5.4.

![Figure 5.4: An alternating series is a sequence of additions and subtractions, which can be visualized as movements left and right on a number line. The terms of the series ($a_n$) are the sizes of the movements, and the partial sums ($S_n$) are the results of consecutive movements. The series depicted has the terms $a_n = (-1)^{n+1}/n$.](image)

The figure makes it quite apparent that the series it depicts converges. The partial sums always have a left bound and a right bound. The left bound starts at 0 and then changes to $S_2$, $S_4$, $S_6$, and so on. The right bound starts at $S_1$, then changes to $S_3$, $S_5$, and so on. At any given point (i.e., vertical position on the diagram), all future partial sums are between the two current bounds. Furthermore, both bounds continue to move toward each other. If the distance between the bounds shrinks to zero as $n$ increases without bound, then surely the series must converge, because the sequence of partial sums would have to converge to the meeting point of the left and right bounds. (This is the squeeze theorem, sort of.)
What we would like to do now is devise precise conditions that a series must obey such that the discussion above is applicable and thus that the series converges.

1. The bounds must always be moving toward one another as \( n \) increases. We can see in the figure that the right bound, for instance, always moves to the left. For instance, \( S_3 \) is farther to the left than \( S_1 \). This is because the change from \( S_1 \) to \( S_3 \) is negative, which is only true because \( a_2 \) (a move to the left) is greater than \( a_3 \) (a move to the right). Similarly, the left bound will only continue to move to the right if \( a_4 \) is less than \( a_3 \), \( a_6 \) is less than \( a_5 \), and so on. This gives us the condition that the sequence of terms \( a_n \) must be decreasing (or at least nonincreasing).

2. Also, the distance between the bounds must eventually converge to zero. Since at any given point the distance between the bounds is equal to a term of the series, this condition is equivalent to the condition that \( \lim_{n \to \infty} a_n = 0 \).

From this, we get:

**Theorem 5.** If the magnitudes of the terms of an alternating series are nonincreasing, and their limiting value is zero, then the series converges.

Figure 5.4 also illustrates an interesting and very useful property of alternating series. Suppose we have calculated the the fourth partial sum, \( S_4 \), and want to use it to estimate the sum of the entire series. For a general series, there is no way to know how good this approximation is. However, alternating series have left and right bounds on their partial sums, as we used in the “proof” of the theorem above. In particular, we know that all partial sums \( S_n \) lie between \( S_4 \) and \( S_5 \), where \( n > 5 \). This means that the sum \( S \) of the entire series must also lie between \( S_4 \) and \( S_5 \). So if we estimate \( S \) by \( S_4 \), then the error cannot be more than the distance between \( S_4 \) and \( S_5 \), which is \( a_5 \). Symbolically, the error in estimating \( S \) by \( S_4 \) is \( |S - S_4| \), and so we have that \( |S - S_4| \leq a_5 \). This same reasoning will apply for any \( S_n \), and so we find that:

**Theorem 6.** For any alternating series whose terms decrease in magnitude monotonically to zero, \( |S - S_n| \leq a_{n+1} \).

### 5.5 The Ratio and Root Tests

**Prerequisite Knowledge.** Understand the geometric series test and the direct comparison test.

The principal idea of the ratio and root tests is to classify a series’ vergence based on how fast it decreases (or increases). For a general series \( \sum_{n=1}^{\infty} a_n \), both \( |a_{n+1}/a_n| \) and \( \sqrt[n]{a_n} \) give quantitative measures of its rate of decrease (for which they are less than 1) or increase (for which they are greater than 1). We will discuss the ratio test first, using figure 5.5 to see graphically how a series’ common ratio affects its convergence or divergence. Take a moment now to study the figure (on page 37). It is complex, and so we will discuss each aspect of it in detail.

Perhaps the easiest series to analyze are geometric series, which have a constant common ratio. For the geometric series \( \sum_{n=0}^{\infty} ar^n \), convergence occurs for \( |r| < 1 \) and divergence occurs for \( |r| \geq 1 \) (according to the geometric series test). Three geometric series are shown in figure 5.5: those with terms \((1/4)^n\), \(1\), and \((3/2)^n\). The common ratios of these series are \(1/4\), \(1\), and \(3/2\) respectively; since these are constants, they appear as horizontal lines when plotted as functions of \( n \). All
Figure 5.5: The common ratios $R$ of several series plotted as functions of $n$. Each plot is labeled with the $n$th term of its series ($a_n$) and, after the tilde ($\sim$), the series’ common ratio ($R_n = a_{n+1}/a_n$). The ratio test can be used to show that all the series in green converge and all the series in red diverge; the test cannot be applied to any of the series in orange. The boundary between convergence and divergence is shown in black. The series corresponding to the sinusoid, which is not labeled on the diagram due to its size, is $\sum_{n=1}^{\infty} \prod_{k=1}^{n} [0.95 + \frac{1}{10} \sin k]$, and its common ratio is $0.95 + \frac{1}{10} \sin n$. (Note: the slope of the $n!$ series has been reduced so that it does not leave the page.)
geometric series may be placed on figure 5.5. If they were, there would be a block of solid green below \( R = 1 \) and a block of solid red above \( R = 1 \), with the series \( \sum_{n=1}^{\infty} 1 \) acting as the boundary. Notice that the other series in the diagram follow a similar pattern: ones mostly below \( R = 1 \) are convergent and ones mostly above \( R = 1 \) are divergent, with a few exceptions.

Of course, there are also geometric (and other) series with negative common ratios, and these are dealt with the same way as series with positive common ratios: figure 5.5 and its coloring are completely symmetric with respect to the \( n \)-axis because the geometric series test depends on \(|r|\), not \( r \).

Now let us turn to the non-geometric convergent series in figure 5.5. First consider the one with terms \( 1/n! \). Past \( n = 3 \), its common ratio is less than that of the convergent geometric series \((1/4)^n\). This means that we can construct another geometric series whose fourth term is equal to the fourth term of \(1/n!\) and whose common ratio is equal to that of \((1/4)^n\). We can call this series \( g_n \). If you’re in the market for a precise formula, it is \( g_n = (1/4!)(1/4)^{n-4} \). Since the common ratio \((R)\) of \(1/n!\) is less than that of \( g_n \) at \( n = 4 \), and since \( a_5 = a_4 \times R_4 \) for both series, we know that the fifth term of \(1/n!\) is less than that of \( g_n \). And since the common ratio of \(1/n!\) continues to be less than that of \( g_n \) for all \( n > 4 \), we can conclude that all terms of \(1/n!\) are less than the corresponding terms of \( g_n \). According to the direct comparison test, this means that \(1/n!\) is convergent.

So what? Isn’t it obvious, one might ask, that \(1/n!\) converges? Perhaps yes, but our argument above using the common ratios of a given, non-geometric series and a different, geometric series can be extended easily. In figure 5.5, the argument translates to: given an arbitrary series, plot its common ratio as a function of \( n \). If we can draw a horizontal line below \( R = 1 \) (corresponding to a convergent geometric series) such that at some point the graph of the series’ common ratio dips permanently below the horizontal line (corresponding to the series’ terms becoming permanently smaller than the convergent geometric series), then we can conclude that the series is convergent.

Now take a moment to consider the two other convergent series in figure 5.5. Can you use the reasoning above to conclude that they are convergent, even though their common ratios do not decrease to zero like that of \(1/n!\)? [You can check the horizontal asymptotes of the graphs by reading the equations for the common ratios, which are labeled on each graph to the right of the tilde (\( \sim \)).]

The divergent series in figure 5.5 can be analyzed in a similar fashion to the convergent ones. You can always use the divergent geometric series with terms \( a_n = 1 \) and common ratio \( R = 1 \): if a series’ common ratio is always larger than 1, then by the direct comparison test it must diverge. (Can you see how to construct a rigorous argument as we did above?)

What about the series in orange? Do they converge or diverge? Consider those with terms \(1/n\) and \(1/n^2\). For either series, can we draw a horizontal line below \( R = 1 \) but also above the series’ graph? It may look like we can, but if you check the formulae for each on the graph, you can see that \( \lim_{n \to \infty} R_n = 1 \) for both. In other words, both series get arbitrarily close to \( R = 1 \), and so it is impossible to show that they converge by drawing a line that is strictly between them and \( R = 1 \). Since both series’ graphs are entirely below \( R = 1 \), we also cannot show that they diverge by showing that they are above \( R = 1 \). So we cannot apply the reasoning shown above to either series. In fact, \(1/n\) diverges and \(1/n^2\) converges. So there is still a border between the convergent and divergent series we can identify using common ratios, where some series converge and some series diverge without any apparent pattern.

The sinusoid is a third example of a series whose vergence cannot be determined by comparison with geometric series. Clearly it passes above any line below \( R = 1 \), but it also passes below any line above \( R = 1 \). Therefore we cannot compare it to a geometric series using common ratios. (This series converges.)
Hopefully, the intuitive idea of determining series’ vergence by their common ratios is now natural to you. We will now translate our criteria above into an easier-to-work-with definition. We have two criteria:

1. \( R_n \) is greater than 1 for all sufficiently large \( n \). This is equivalent to the statement \( \lim_{n \to \infty} R_n > 1 \). \(^5\) (Think about this until you are convinced. You can check that the two statements are equivalent for each of the divergent series in figure 5.5.) From this, we have that if \( \lim_{n \to \infty} (a_{n+1}/a_n) > 1 \), then \( \sum a_n \) diverges.

2. There is a number \( L \), less than 1, such that \( R_n \) is less than \( L \) for all sufficiently large \( n \). This is equivalent to the statement \( \lim_{n \to \infty} R_n < 1 \). [Also convince yourself of this. You can check the equivalence of the statements for both the convergent (where they are both true) and ambiguous (where they are both false) series labeled in the figure.] From this, we have that if \( \lim_{n \to \infty} (a_{n+1}/a_n) < 1 \), then \( \sum a_n \) converges.

Since, as we mentioned earlier, all of this reasoning is symmetric with respect to the \( n \)-axis (only series with positive common ratios were discussed for the sake of brevity), we should really replace \( a_{n+1}/a_n \) with \( |a_{n+1}/a_n| \) to be more general. Thus, we obtain what is commonly known as the ratio test:

**Theorem 7.** Let \( L = \lim_{n \to \infty} |a_{n+1}/a_n| \). If \( L > 1 \), then \( \sum_{n=1}^{\infty} a_n \) diverges; and if \( L < 1 \), then \( \sum_{n=1}^{\infty} a_n \) converges. If \( L = 1 \) or the limit does not exist, then no conclusion can be drawn.

On a broad scale, let’s summarize what we have done. We have identified a property,

\[
R_n = \left| \frac{a_{n+1}}{a_n} \right|,
\]

which is constant for geometric series \( R_n = |r| \) for \( \sum r^n \) and can be used to determine the vergence of those series: \( R_n < 1 \) means convergence; \( R_n \geq 1 \) means divergence), then extended it to other series in the limit, except with the boundary line being ambiguous: \( \lim_{n \to \infty} R_n < 1 \) means convergence, \( \lim_{n \to \infty} R_n > 1 \) means divergence, and \( \lim_{n \to \infty} R_n = 1 \) is inconclusive.

There is, in fact, another property \( R'_n \) which is also constant and equal to \( |r| \) for geometric series \( \sum r^n \). It is \( \sqrt[n]{|a_n|} \). We could go through a rigorous proof, but what is the fun in that? We can instead simply note that our new value \( R'_n \) has the same property for geometric series as \( R \) did, and realize that it will probably naturally generalize in the same way (it does): that is, \( \lim_{n \to \infty} R'_n < 1 \) means convergence, \( \lim_{n \to \infty} R'_n > 1 \) means divergence, and \( \lim_{n \to \infty} R'_n = 1 \) is inconclusive. Thus, we have the root test:

**Theorem 8.** Let \( L = \lim_{n \to \infty} \sqrt[n]{|a_n|} \). If \( L > 1 \), then \( \sum_{n=1}^{\infty} a_n \) diverges; and if \( L < 1 \), then \( \sum_{n=1}^{\infty} a_n \) converges. If \( L = 1 \) or the limit does not exist, then no conclusion can be drawn.

As an addendum, here are four useful limits which make it much easier to apply the root test to a wide class of functions:

\[
\lim_{n \to \infty} \sqrt[n]{C} = 1,
\]

\(^5\)Technically, this is not quite true. In some cases, the first statement can be true even though the limit does not exist, such as if \( R_n = \frac{1}{n}(1 + \sin n) \). However, if the limit exists, the first statement always has to be true. To make the two statements equivalent, we would need to use a limit inferior instead of a regular limit. See [https://en.wikipedia.org/wiki/Limit_inferior](https://en.wikipedia.org/wiki/Limit_inferior) for details if you are willing to venture out into that untamed wilderness.
\[ \lim_{n \to \infty} \sqrt[n]{n} = 1, \]
\[ \lim_{n \to \infty} \sqrt[n]{n!} = \infty, \]
and
\[ \lim_{n \to \infty} \frac{\sqrt[n]{n!}}{n} = \frac{1}{e}. \]

The first two limits may be proven rather easily using logarithms, and proofs of the last two can be found at http://matharticles.com/ma/ma050.pdf.

6 Vector Analysis

6.1 The Gradient

**Prerequisite Knowledge.** Understand the idea of scalar and vector fields in two and three dimensions, approximating a function with its tangent plane, partial and directional derivatives, and projections of vectors.

6.1.1 Intuition

The gradient is an operation that takes a scalar field \( f \) and returns a vector field, \( \nabla f \). The best way to understand the gradient is to look at a diagram such as figure 6.1, which shows a two-dimensional scalar field and its gradient.

The gradient at a particular point is a vector pointing in the direction of greatest increase and with a magnitude equal to that rate of greatest increase. It should make sense that this idea is still perfectly valid in three dimensions even though figure 6.1 only shows a gradient in two dimensions.

6.1.2 Definition

It turns out that the gradient and directional derivatives of a function are intimately related. In fact, given just the gradient at a single point, you can work out all of the directional derivatives at that point! How, you ask?

Let’s work with a function of two variables. To make things simpler, instead of looking directly at the three-dimensional graph of this function \( z = f(x, y) \), let’s just work with its tangent plane (which will exist if and only if the function is differentiable). Just as the derivative of a function is the same as the slope of its tangent line, the directional derivative of a function is the same as the slope of its tangent line in the same direction. This tangent line will, of course, be contained within the tangent plane. In figure 6.3, the blue plane is the tangent plane to a function at the indicated point. Figure 6.3 also shows several tangent lines to the same function, and the slope of each one is equal to the directional derivative in the direction of the tangent line.

In particular, the slope of the vector marked “direction of greatest increase” is equal to the greatest possible directional derivative, the slope of the vector marked “direction of greatest decrease” is equal to the least possible directional derivative, the slope of the vector marked “direction of zero increase” is (of course) zero, and the slope of the vector marked “arbitrary direction” is somewhere in between.
Figure 6.1: The graph of a two-dimensional scalar field, where red areas have large values and blue areas have small values. The gradient of this field is overlaid as a grid of arrows. At each point in the field, the gradient is in the direction of greatest increase, and it is larger when that rate of increase is faster. See figure 6.2 for a three-dimensional plot of this scalar field.

There is one more interesting fact that you should be able to convince yourself of if you look at figure 6.3 long enough. This is that the point marked on the figure and the direction of greatest increase uniquely define the blue plane. In other words, it is not possible to draw any other plane that goes through the marked point and has the same direction of greatest increase as the blue plane. What this means is that given just the direction and magnitude of greatest increase (in other words, the gradient), we know the tangent plane and therefore also every directional derivative. But how can we find an explicit formula for any given directional derivative in terms of the gradient?

It’s time for some good old-fashioned coordinate geometry. Using figure 6.4, we can associate several of the quantities we are discussing with geometric properties. On the figure we have drawn two right triangles: one in the direction of maximum increase ($\triangle OAB$) and one in an arbitrary other direction that differs from the first by an angle $\theta$ ($\triangle OCD$). If we draw these triangles such that their bases ($OA$ and $OC$) both have lengths of 1, then the slopes of the hypotenuses will be the lengths of $AB$ and $CD$ respectively. First, we can use the angle $\phi$ to do some trigonometry on $\triangle OPC$. In particular, we have $\sin \phi = PC/OC = PC$ because $OC$ is of length 1. Since the angles $\phi$ and $\theta$ are complementary, we have $\sin \phi = \cos \theta$ and so $PC = \cos \theta$. Now note that $OB$ and $PD$ are parallel, and so their angles $\angle AOB$ and $\angle CPD$ are congruent: both are equal to the angle of elevation of the blue plane relative to the horizontal green plane. Since $\triangle OAB$ and
Figure 6.2: The scalar field displayed in figure 6.1 graphed as a function of two variables. The same colors, representing values of the function, are used here; and grid lines have been added to make the shape of the surface more evident. Note that the arrows in figure 6.1 are largest at the steepest points on this graph, and always point directly up the slope.

\( \triangle PCD \) are right triangles with one congruent angle, all of their angles are congruent and so the triangles are similar. This means that we can equate the ratios \( OA/AB = PC/CD \), or in other words \( 1/AB = \cos \theta/CD \), where we have used the facts from earlier that \( OA = 1 \) and \( PC = \cos \theta \). We can rewrite this equation to see that \( CD = AB \cos \theta \).

Now all that is left is to translate this geometric property back into the language of vectors. Recall that \( AB \) is the slope of \( OB \), which is the rate of maximum increase, which is the magnitude of the gradient. Recall also that \( CD \) is the slope of \( OD \), which is the rate of increase in the direction \( \theta \) away from the direction of maximum increase, which is the directional derivative in this direction. We then have \( D_u f = |\text{grad} f| \cos \theta \), where \( D_u f \) is the directional derivative in the direction of \( u \) and \( u \) is a unit vector in the direction of \( OD \). There you have it!

But there is an even more convenient form of this equation. Recall the geometric formula for the dot product, \( \mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos \theta \). If \( \mathbf{b} \) is a unit vector \( u \), then we have \( |u| = 1 \) and so \( \mathbf{a} \cdot \mathbf{u} = |\mathbf{a}| \cos \theta \). Wow – that looks familiar! All we have to do is let \( \mathbf{a} = \text{grad} f \) and we have that \( \text{grad} f \cdot \mathbf{u} = |\text{grad} f| \cos \theta = D_u f \). So we have proved that the dot product of the gradient with any unit vector is the directional derivative in that direction.

Because the component of \( \mathbf{v} \) in the direction of the unit vector \( \mathbf{u} \) is \( \mathbf{v} \cdot \mathbf{u} \), we need only let \( \mathbf{v} = \text{grad} f \) to see that the component of the gradient in any direction is the directional derivative in that direction. Thus, we have arrived at another, subtler, intuitive interpretation of the gradient!

Typically, the gradient is defined as the unique vector whose component in any given direction is the directional derivative in that direction. I, however, think that it is more natural to start with the definition of the gradient as the direction and magnitude of maximum increase and derive the
Figure 6.3: Various vectors pointing away from a specific point on a tangent plane (the blue plane), with a horizontal plane (the green plane) provided for reference. The slope of any given vector corresponds to the directional derivative in the direction of the vector. In particular, the vector marked “direction of greatest increase” corresponds to the maximum possible directional derivative, and as such is equal to the gradient of the function.

Figure 6.4: Two directional derivatives differing by an angle \( \theta \), where one is in the direction of maximum increase, are visualized here as line segments with slopes equal to the corresponding derivatives. A second right triangle parallel to \( \triangle OAB \) has been added to the other elements present in figure 6.3.
6.1.3 Definition with Partial Derivatives

The gradient also has a formula in terms of partial derivatives. It is quite easy, really: if we want to express the gradient in terms of \( \mathbf{i}, \mathbf{j}, \) and \( \mathbf{k} \), all we need are the components of the gradient in those three directions. We already showed that the component of the gradient in any given direction is simply the directional derivative in that direction. In fact, the directional derivatives in the \( \mathbf{i}, \mathbf{j}, \) and \( \mathbf{k} \) directions already have names: they are better known as the partial derivatives with respect to \( x, y, \) and \( z \). So the gradient is also given by:

\[
\nabla f = \frac{\partial f}{\partial x} \mathbf{i} + \frac{\partial f}{\partial y} \mathbf{j} + \frac{\partial f}{\partial z} \mathbf{k}.
\]

The nice, symmetric form taken by this last form of the gradient has inspired mathematicians to write the equation in the following form:

\[
\nabla f = \left( \frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k} \right) f.
\]

If you pretend, for a moment, that a vector made up of partial derivative operators is even a legitimate mathematical structure (of course, it is not, but bear with me for a moment here), and simply multiply the “vector” by the “scalar” \( f \), then you will get the previous formula for the gradient in terms of partial derivatives. For convenience, many people like to call the “vector” of partial derivative operators by the name “del” or “nabla”, written as \( \nabla \) (you may see this symbol as \( \nabla \) but I think it should be in \textbf{boldface} because it is a vector). These people then like to write the gradient as

\[
\nabla f = \nabla f.
\]

I think it’s a little confusing, but you can write it however you like. Any good mathematician (or student) should understand either.

6.2 The Gradient Theorem

**Prerequisite Knowledge.** Understand the gradient (see section 6.1), the proof of the first fundamental theorem of calculus (see section 4.2), and line integrals.

6.2.1 The Equation

The gradient theorem is a generalization of the first fundamental theorem of calculus to a curved path rather than a straight line. Recall that the fundamental theorem is calculus is:

\[
\int_a^b f'(x) \, dx = f(b) - f(a).
\]

The gradient theorem is:

\[
\int_a^b \nabla f(x) \cdot \, dx = f(b) - f(a).
\]

In this equation, \( f \) is a function of multiple variables, but rather than writing \( f(x, y) \) we just let \( \mathbf{x} = x \mathbf{i} + y \mathbf{j} \) and say \( f(\mathbf{x}) \). You can see that although the left-hand side is a line (or path, or contour) integral, the path is not given. In fact, according to the gradient theorem, the integral is the same value, \( f(b) - f(a) \), no matter what path is used!

You may also see the gradient theorem written in the following equivalent form:

\[
\int_a^b \nabla f \cdot \, dx = f(b) - f(a).
\]
6.2.2 The Proof

In any case, the important thing is understanding why the theorem is true. You may wish to review section 4.2 – and in particular 4.2.4 – because the proof of the gradient theorem is exactly analogous to that of the first fundamental theorem of calculus.

<table>
<thead>
<tr>
<th>First fundamental theorem of calculus</th>
<th>Gradient theorem in two dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The total change in $y$ is $\Delta y = f(b) - f(a)$.</td>
<td>The total change in $z$ is $\Delta z = f(b) - f(a)$.</td>
</tr>
<tr>
<td>The interval from $a$ to $b$ can be split into many subintervals, each of which has a $\Delta x_i$. Over each subinterval, the value of $f(x)$ will change by $\Delta y_i$.</td>
<td>The curve from $a$ to $b$ can be split into many small sections of curve, each of which has a $\Delta x_i$ and $\Delta y_i$. Over each section of curve, the value of $f(x)$ will change by $\Delta z_i$.</td>
</tr>
<tr>
<td>We can approximate $\Delta y_i$ using the tangent line to the function.</td>
<td>We can approximate $\Delta z_i$ using the tangent plane to the function.</td>
</tr>
<tr>
<td>The tangent line has a slope of $\frac{df}{dx}$, so changing $x$ by $\Delta x_i$ will change $y$ by $\frac{df}{dx} \Delta x_i$.</td>
<td>We can split the change in $z$ into two parts: the part due to changing $x$ and the part due to changing $y$. The tangent plane has a slope of $\frac{\partial f}{\partial x}$ in the $x$ direction, so changing $x$ by $\Delta x_i$ will change $z$ by $\frac{\partial f}{\partial x} \Delta x_i$. On the other hand, the tangent plane has a slope of $\frac{\partial f}{\partial y}$ in the $y$ direction, so changing $y$ by $\Delta y_i$ will change $z$ by $\frac{\partial f}{\partial y} \Delta y_i$. Since we change both $x$ and $y$, we can simply add the two resulting changes in $z$ together to get the total change in $z$.</td>
</tr>
<tr>
<td>Therefore, $\Delta y_i \approx \frac{df}{dx} \Delta x_i$.</td>
<td>Therefore, $\Delta z_i \approx \frac{\partial f}{\partial x} \Delta x_i + \frac{\partial f}{\partial y} \Delta y_i$.</td>
</tr>
</tbody>
</table>
Since adding together all of the $\Delta y_i$ will give us the total $\Delta y = f(b) - f(a)$, we have

$$f(b) - f(a) = \sum_i \Delta y_i$$

$$\approx \sum_i \frac{df}{dx} \Delta x_i$$

$$\approx \sum_i f'(x) \Delta x_i$$

$$= \int_a^b f'(x) \, dx.$$

Since adding together all of the $\Delta z_i$ will give us the total $\Delta z = f(b) - f(a)$, we have

$$f(b) - f(a) = \sum_i \Delta z_i$$

$$\approx \sum_i \frac{\partial f}{\partial x} \Delta x_i + \frac{\partial f}{\partial y} \Delta y_i$$

$$\approx \sum_i \left( \frac{\partial f}{\partial x} \mathbf{i} + \frac{\partial f}{\partial y} \mathbf{j} \right) \cdot (\Delta x_i \mathbf{i} + \Delta y_i \mathbf{j})$$

$$\approx \sum_i (\text{grad } f) \cdot (\Delta x_i)$$

$$= \int_a^b \text{grad } f \cdot dx.$$

Although this proof was for functions of two variables, it is quite easy to extend it to functions of three variables. In general, the gradient theorem is true for any number of dimensions.

Also, note that we made no reference in the proof to the overall shape of the path of the line integral; thus, the path must not affect the value of the integral!

### 6.3 The Divergence

**Prerequisite Knowledge.** Understand line integrals, surface integrals, their interpretations as flux, and partial derivatives.

#### 6.3.1 Intuition

The divergence is an operation that takes a vector field $\mathbf{F}$ and returns a scalar field, $\text{div } \mathbf{F}$. Be careful, though, that you do not think the divergence and gradient are inverse functions!

The most intuitive way to visualize the divergence is with a velocity field. Given a region (two-dimensional or three-dimensional) filled with moving liquid, at each point the liquid is moving a certain velocity, which is a vector (direction and magnitude). Four velocity fields are pictured in figure 6.5.

The divergence of a velocity field at a particular point represents the expansion (positive), compression (negative), or lack of either (zero) of the fluid flow at that point.

For instance, the top-left field in figure 6.5 shows a fluid with constant velocity. At any given point, the velocity of fluid flowing in is the same as the velocity of fluid flowing out. Everything balances, there is no compression or expansion, and there is no divergence.

In contrast, the top-right field in figure 6.5 shows a fluid that is accelerating. Since the velocity is greater farther away from the origin, at any given point the fluid flowing in from the origin will be moving slower than the fluid flowing out away from the origin. Since the outflow is greater than the inflow, more fluid must somehow be being added in order to make up the difference. If we assume that the conservation of mass precludes fluid appearing out of thin air, then the only way for less fluid to turn into more fluid is if it expands. (In real life, this might be due to a change in
temperature or pressure. We don’t really care what caused the expansion, though – just the fact that it is happening.) Thus, the divergence is positive everywhere in the top-right field. (In fact, the divergence is the same everywhere in that field. The important part is change in velocity, not just a field that “looks” like it is expanding, as is true near the center.)

Velocity fields can have zero divergence (be “divergence free”) even if they are not constant. For instance, the bottom-left field in figure 6.5 has no divergence anywhere. You might wonder how the velocity can flow toward the origin along the $x$-axis without any compression, but note that unless the fluid is precisely on the $x$-axis it will break away and move up or down. The fluid that is precisely on the $x$-axis has an infinitely small volume in any case, so it does not need to be considered. Also, as the fluid moves up the $y$-axis its velocity increases, but this effect is counterbalanced by more fluid moving toward the $y$-axis, thus making the existing fluid flow faster. Overall, everything cancels out and there is no divergence. (We’ll see the mathematics behind this momentarily – it’s not just visual!)

The bottom-right field in figure 6.5 is somewhat counterintuitive: it looks like the fluid is
expanding because it is moving away from the origin, but in fact its divergence is negative! Why? Although the fluid is flowing outward, it is also slowing down very quickly. This means that the amount of fluid flowing out of any given point is less than the amount of fluid flowing into it. In general, it’s quite difficult to tell what the divergence of a field is just from a picture. (In particular, if you don’t know the magnitudes of the vectors, there’s no way at all to tell!) A radial velocity field can have positive, negative, or zero divergence.

We have been discussing velocity fields in two dimensions, but of course the same ideas apply for vector fields that do not represent the velocity of a fluid and for vector fields in three or more dimensions.

6.3.2 Definition

In our intuitive discussion, we mentioned that divergence – positive or negative – occurs when there is an imbalance in inflow and outflow in a particular region. Inflow and outflow can be quantified using the concept of flux, which is defined based on either a line integral or a surface integral depending on the number of dimensions. In two dimensions, the net flux of $\mathbf{F}$ through a curve $C$ is

$$\int_C \mathbf{F} \cdot d\mathbf{s},$$

and in three dimensions, the net flux of $\mathbf{F}$ through a surface $S$ is

$$\iint_S \mathbf{F} \cdot d\mathbf{S}.$$

In the first formula, $d\mathbf{s}$ is a normal vector to the curve, pointing in the direction in which we are calculating flux and with a magnitude corresponding to its length $ds$ on the curve. In the second, $d\mathbf{S}$ is a normal vector to the surface, pointing in the direction in which we are calculating flux and with a magnitude corresponding to its surface area $dS$ on the surface.

We will be investigating flux out of closed curves and surfaces in our definition of divergence. (Positive divergence will correspond to net flux out and negative divergence will correspond to net flux in.) However, the flux out of any curve or surface will depend on the value of the function at every point on the curve or surface.

When we defined the derivative, we started with a non-local property, the slope of a secant line. This property depended on points significantly far away from the point of interest. To obtain the local property of the slope of the tangent line – which depends only on points very, very (arbitrarily) close to the point of interest – we had to take a limit. Similarly, the flux through a curve or surface is a non-local property, and we will have to take a limit in order to obtain a local property – the divergence, which should only depend on points very, very close to the point of interest. For instance, in the two-dimensional case we could try

$$\lim_{\text{Area}(C) \to 0} \int_C \mathbf{F} \cdot d\mathbf{s},$$

where Area$(C)$ is the area enclosed by $C$ (remember that $C$ is a closed curve). Unfortunately, this limit is always zero because as the area enclosed by the curve gets smaller, so will the net flux in or out of it.

When we defined the derivative, we used the difference $f(x) - f(c)$. However, just taking the limit as $x \to c$ would make this difference go to zero, so we had to add a balancing factor of $x - c$ to the denominator. Similarly, here we will have to add a balancing factor of Area$(C)$ to the
denominator. So a reasonable definition giving the net amount of outward flux as a characteristic property of a particular point is
\[
\text{div } \mathbf{F} = \lim_{\text{Area}(C) \to 0} \frac{\int_C \mathbf{F} \cdot ds}{\text{Area}(C)}.
\]
In three dimensions, this would be
\[
\text{div } \mathbf{F} = \lim_{\text{Volume}(S) \to 0} \frac{\iint_S \mathbf{F} \cdot dS}{\text{Volume}(S)},
\]
and the idea generalizes nicely to higher dimensions. These limits give rise to the informal definitions of divergence as “flux per unit area” or “flux per unit volume”, just as the derivative is the “change in \( y \) per unit change in \( x \)”. You may have wondered why we did not use arc length instead of enclosed area, or surface area instead of enclosed volume. The problem with these quantities is that they can be changed immensely without changing the amount of net flux. You can see how to increase the arc length of a contour without changing its net flux in figure 6.6.

![Figure 6.6: The black contour represents a region which will have a certain amount of net flux. If the gray contour is added, the arc length of the contour will be increased greatly, but the flux will not change significantly since any flux in the region of the gray contour will just enter one side and exit the other – the vector field does not have the chance to change significantly across the width of the gray contour.](image)

Of course, we didn’t give a specific reason why the formula takes the particular form it does, but then again defining “amount of net flux as a function of position” is a far more nebulous proposition than defining “slope of the tangent line as a function of position”, which is quite straightforward. In the end, the reasonability of this definition of divergence is tested by seeing if its behavior agrees with what we would intuitively expect.

You will see in the next section that this definition gives a very nice, concise formula for the divergence of a field.

### 6.3.3 Definition with Partial Derivatives

Let’s derive a more straightforward formula for the divergence in two dimensions. To evaluate the limit in the definition, we’ll consider a contour that is so small we can calculate the line integral directly. Such a contour is shown in figure 6.7.
The contour used in the definition can be approximated as a rectangle. The proof we are about to follow will work for any contour, but this simplified version should make the idea clear. We want to find

$$\frac{\int_C \mathbf{F} \cdot d\mathbf{s}}{\text{Area}(C)}.$$  

The area enclosed by $C$ is easy to find using the coordinates on the figure: it is simply $\Delta x \Delta y$. To find the net flux out of $C$, we can look at each of the four sides individually and add their flux together.

First, take the bottom side. We can assume that $\mathbf{F}$ is relatively constant over such a short range ($\Delta x$), and say it has the value of $\mathbf{F}(x, y)$ everywhere on the bottom side. Then, since $\mathbf{F}$ is constant, we can compute the line integral

$$\int_{C_{\text{bottom}}} \mathbf{F}(x, y) \cdot d\mathbf{s}$$

directly. Remember that $d\mathbf{s}$ is a normal vector pointing out of the curve; for the bottom side we will have $d\mathbf{s} = -\mathbf{j} \, dx$. We will thus have

$$\int_{C_{\text{bottom}}} (F_x(x, y) \mathbf{i} + F_y(x, y) \mathbf{j}) \cdot (-\mathbf{j} \, dx) = \int_x^{x+\Delta x} -F_y(x, y) \, dx = -F_y(x, y)\Delta x.$$

You can (and should) compute the fluxes for the right, top, and left sides as well. They are $F_x(x+\Delta x, y)\Delta y$, $F_y(x, y+\Delta y)\Delta x$, and $-F_x(x, y)\Delta y$ respectively. We can find the net flux out of the contour by summing all four of these numbers together:

$$\int_C \mathbf{F} \cdot d\mathbf{s} = F_x(x+\Delta x, y)\Delta y - F_x(x, y)\Delta y + F_y(x, y+\Delta y)\Delta x - F_y(x, y)\Delta x.$$  

Then, we can simply apply the definition:

$$\text{div} \mathbf{F} = \lim_{\Delta x, \Delta y \to 0} \frac{\int_C \mathbf{F} \cdot d\mathbf{s}}{\text{Area}(C)}$$

$$= \lim_{\Delta x, \Delta y \to 0} \frac{F_x(x+\Delta x, y)\Delta y - F_x(x, y)\Delta y + F_y(x, y+\Delta y)\Delta x - F_y(x, y)\Delta x}{\Delta x \Delta y}$$

$$= \lim_{\Delta x, \Delta y \to 0} \left[ \frac{F_x(x+\Delta x, y)\Delta y}{\Delta x \Delta y} + \frac{F_y(x, y+\Delta y)\Delta x}{\Delta x \Delta y} - \frac{F_y(x, y)\Delta x}{\Delta x \Delta y} \right].$$

Figure 6.7: A small contour that can be used in the definition of divergence (and curl; see page 54). The rectangle has a width of $\Delta x$ and a height of $\Delta y$. We can approximate $\mathbf{F}$ as constant along each side, although a fully rigorous proof would not do so. The flux out of each side depends on the length of the side and the component of $\mathbf{F}$ normal to the side.
In the last step, we have simply used the definition of partial derivative. You can do the same proof except with a rectangular prism instead of a rectangle, and you will see that the divergence in three dimensions is exactly analogous to the divergence in two dimensions:

\[
\text{div } \mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}.
\]

Recalling the symbolic convenience

\[
\nabla = \frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k},
\]

you can also write this as

\[
\text{div } \mathbf{F} = \nabla \cdot \mathbf{F}.
\]

6.4 The Divergence Theorem

**Prerequisite Knowledge.** Understand divergence (see section 6.3), flux, and volume integrals.

The divergence theorem is a natural consequence of the definition of divergence, expanded to a macroscopic scale (rather than the microscopic scale used in the definition). In two dimensions, it states that

\[
\iint_R \text{div } \mathbf{F} \, dA = \oint_C \mathbf{F} \cdot d\mathbf{s},
\]

where \( C \) is a closed contour that bounds the region \( R \). In three dimensions, it states that

\[
\iiint_Q \text{div } \mathbf{F} \, dV = \iint_S \mathbf{F} \cdot d\mathbf{S},
\]

where \( S \) is a closed surface that bounds the region \( Q \). Here, \( \text{div } \mathbf{F} \) should be continuous. (There are some odd cases of vector fields with non-continuous divergence that would otherwise violate the divergence theorem.)

We will prove the two-dimensional case. First consider a region that is split into two subregions, as shown in figure 6.8.

What is the net flux out of the region? It is the sum of the net fluxes out of \( A \) and \( B \). All flux that exits \( A \) through the vertical line will just enter \( B \), and vice versa. The flux along the vertical line has the same magnitude but opposite sign in \( A \) and \( B \), so when the two fluxes are added together this component will cancel out. All that is left is the flux along the boundary of the region, which means that the sum of the net fluxes out of \( A \) and \( B \) gives the flux out of the total region they make up.

In general, a region can be divided into any number of subregions, and the total flux out of the region can be found by adding together the net fluxes out of each subregion. In addition, this reasoning will also apply to three dimensions, where a region in space will be divided into subregions divided by surfaces.
Figure 6.8: An arbitrary region $R$ divided into two subregions $A$ and $B$. The boundary line between the regions does not add any flux since the flux out of $A$ through the boundary will be exactly the same as the flux out of $B$ through the boundary, but with opposite sign.

How does this help us prove the divergence theorem? Recall that the definition of divergence is "net flux out divided by area, for very small area". This could be written as

$$\text{div } F = \frac{d(\text{net flux out})}{dA}.$$ 

To find the total flux out of a large number of these small regions, we must multiply each of them by $dA$ and then add them all together. [In other words, $\iint_R d(\text{net flux out}) = \text{total flux out.}$] In fact, this is exactly what the divergence theorem states:

$$\iint_R \text{div } F \, dA = \int_C F \cdot ds.$$ 

Of course, a similar proof can be followed in three or more dimensions.

### 6.5 The Curl

**Prerequisite Knowledge.** Understand line integrals and partial derivatives.

#### 6.5.1 Intuition

The curl is an operation that takes a vector field $F$ and returns a different vector field, $\text{curl} F$.

Suppose again that $F$ is the velocity field of a fluid (in three dimensions), and suppose that we have placed a very small sphere with a rough surface at the point of interest in this field. Suppose further that we are somehow able to prevent this sphere from moving in any direction but also allow it to rotate in any direction without resistance. (Never mind how!) For some velocity fields, the motion of the fluid combined with the rough surface of the sphere will cause it to rotate. The curl is directly related to the rotation of the sphere.

But we want to be more precise than this. What is the most concise way to describe the rotation of a sphere? The two things that need to be specified are the axis of rotation and the rate of rotation (revolutions per second, or degrees per second, or preferably radians per second). That makes a vector (direction and magnitude)! The only thing missing is specifying which way the rotation is – clockwise or counterclockwise. We will say that the vector shall be oriented in such a way.
that when you look from the direction the vector is pointing to, the sphere appears to be rotating counterclockwise. This setup is illustrated in figure 6.9.

![curl vector](image)

Figure 6.9: The traditional way to describe the rotation of an object (in this case a sphere) using a single vector. This scheme is often referred to as the “right-hand rule”, because if you curl the fingers of your right hand in the direction of rotation of the sphere, then your thumb will point in the direction of the curl (angular velocity) vector.

The rotating sphere analogy gives a nice way to visualize an arbitrary component of the curl. In figure 6.9, consider passing a frictionless vertical rod through the sphere, so that it could only rotate around a vertical axis. Then the sphere will have a new “rotation vector”, which will represent only the rotation about a vertical axis. This vector will be the component of the unconstrained rotation vector in the direction of the vertical axis. We’ll return to this way of finding a component of the curl momentarily.

### 6.5.2 Definition

We saw that the gradient could be defined as the unique vector whose components were given by a particular formula. (In the case of the gradient, that formula was the directional derivative in the direction of the component.) We defined the divergence using the limit of a line integral. In our definition of the curl, we will combine both of these strategies.

In the previous section, we saw that the component of the curl in a particular direction was directly related to the angular velocity (rate of rotation) of a rough-surfaced sphere constrained to rotate about an axis in the direction of the component. All we have to do is quantify this property. For simplicity, we will just consider a cross-section of the sphere on the plane normal to the component in which we are interested, which is a circle \( C \). So given the curve \( C \) and a velocity field \( \mathbf{F} \), we want to find a number corresponding to the tendency of the circle (sphere) to rotate counterclockwise (positive) or clockwise (negative). This will be the magnitude of the component of the curl vector pointing out of the page (see figure 6.10) because we defined the vector as pointing in the direction from which you can look to see the sphere (circle) rotating counterclockwise.

We can quantify the tendency of the fluid flow \( \mathbf{F} \) to make the circle rotate by looking at each point on the circle individually and summing the results (this is an integral). If \( d\mathbf{x} \) is a tangent vector that points in the counterclockwise direction along the circle, as shown in figure 6.10, then the tendency of the circle to rotate counterclockwise is maximized when \( \mathbf{F} \) and \( d\mathbf{x} \) are parallel. The tendency is, of course, minimized when the vectors are antiparallel (pointing in opposite directions). And there is no tendency for the sphere to rotate if \( \mathbf{F} \) and \( d\mathbf{x} \) are orthogonal. To be precise, the tendency to rotate counterclockwise is given by \( d\tau = \mathbf{F} \cdot d\mathbf{x} \), where \( \tau \) stands for tendency to rotate (or torque). The total torque is, of course, given by

\[
\tau = \int_C d\tau = \int_C \mathbf{F} \cdot d\mathbf{x}.
\]
Figure 6.10: The tangent vector $\mathbf{d}x$ and the velocity field $\mathbf{F}$ at one point along a contour. In this case, since $\mathbf{F}$ and $\mathbf{d}x$ are pointing in opposite directions, $\mathbf{F}$ would tend to rotate the contour clockwise – the opposite direction of $\mathbf{d}x$. This idea can be quantified using the dot product $\mathbf{F} \cdot \mathbf{d}x$.

This quantity is called the circulation of $\mathbf{F}$ around $C$.

Note here that the differential vector is a tangent vector to the curve and not a normal vector as it was in the definition of flux. (To be precise, the magnitude of $\mathbf{d}x$ is the arc length of its small section of curve.) I have adopted the convention that $\mathbf{d}s$ and $\mathbf{d}S$ are normal vectors while $\mathbf{d}x$ is a tangent vector. (In most sources, you will see $\mathbf{d}n$ and $\mathbf{d}S$ as normal vectors, while $\mathbf{d}s$ is a tangent vector. I think this is confusing. The important thing, of course, is to know what your equations mean intuitively so that you are never confused by notation!)

Naturally the torque on an object affects its rotation, but so does its mass. Since our circular cross-section is a two-dimensional figure, we’ll assume its mass is proportional to its area. An object with twice the mass is twice as hard to rotate, so the rotation, or curl, of our circle will be inversely proportional to its area. Therefore, we have the following formula:

$$\text{component of } \text{curl } \mathbf{F} = \frac{\int_C \mathbf{F} \cdot \mathbf{d}x}{\text{Area}(C)}.$$

Oops! Our curl depends on the size of the circle. We’ll want to use an arbitrarily small region, just as we did for gradient:

$$\text{component of } \text{curl } \mathbf{F} = \lim_{\text{Area}(C) \to 0} \frac{\int_C \mathbf{F} \cdot \mathbf{d}x}{\text{Area}(C)}.$$

Wow! Look how similar that is to the limit definition of divergence. It should come as no surprise that just as the divergence is sometimes called the flux per unit area, the curl is sometimes called the circulation per unit area.

The most important difference is that curl, being a vector, is more complicated than divergence in that its limit formula only gives one component of the curl.

The curl is defined as the unique vector whose component in any direction is given by the formula above, where $C$ is a contour in the plane normal to the direction of the component. (The line integral should also be oriented in a counterclockwise direction when viewed from the direction in which the component is pointing.)

6.5.3 Definition with Partial Derivatives

The curl is so analogous to the divergence that the derivation of a formula for the curl in terms of partial derivatives is essentially the same as the derivation for divergence, just using a slightly
different line integral. The only major difference is that we will have to find the curl one component at a time, because its limit definition only gives us one component. We will find the \( \mathbf{k} \)-component first. The figure from the divergence proof is reprinted here for your convenience:

\[
\begin{align*}
(x, y + \Delta y) & \quad \mathbf{F} \approx \mathbf{F}(x, y + \Delta y) \quad (x + \Delta x, y + \Delta y) \\
(x, y) & \quad \mathbf{F} \approx \mathbf{F}(x, y) \quad (x + \Delta x, y) \\
(x, y + \Delta y) & \quad \mathbf{F} \approx \mathbf{F}(x, y + \Delta y) \\
(x, y) & \quad \mathbf{F} \approx \mathbf{F}(x, y)
\end{align*}
\]

Notice that in this figure the \( \mathbf{k} \) unit vector is pointing up out of the page, and so the contour pictured is indeed normal to the component we are interested in (the \( \mathbf{k} \)-component). Of course, the line integral should be evaluated in the counterclockwise direction.

This time, we must again evaluate the dot product in the line integral for each of the four sides, except with \( \mathrm{d}x \) instead of \( \mathrm{d}s \) (a tangent vector instead of a normal vector). For instance, for the bottom side, \( \mathrm{d}x = \mathbf{i} \, \mathrm{d}x \) and so

\[
\int_{C_{\text{bottom}}} (F_x(x, y) \mathbf{i} + F_y(x, y) \mathbf{j}) \cdot (\mathbf{i} \, \mathrm{d}x) = \int_x^{x+\Delta x} F_x(x, y) \, \mathrm{d}x = F_x(x, y) \Delta x.
\]

You should evaluate the line integrals for the other three sides. You should get \( F_y(x + \Delta x, y) \Delta y \), \( -F_x(x, y + \Delta y) \Delta x \), and \( -F_y(x, y) \Delta y \) for the right, top, and left sides respectively.

We can find the total circulation of \( \mathbf{F} \) around \( C \) by summing these four components:

\[
\int_C \mathbf{F} \cdot \mathrm{d}x = F_y(x + \Delta x, y) \Delta y - F_y(x, y) \Delta y + F_x(x, y) \Delta x - F_x(x, y + \Delta y) \Delta x.
\]

Then, we can simply apply the definition of the \( \mathbf{k} \)-component of curl:

\[
\text{curl}_k \mathbf{F} = \lim_{\text{Area}(C) \to 0} \frac{\int_C \mathbf{F} \cdot \mathrm{d}x}{\text{Area}(C)}
\]

\[
= \lim_{\Delta x, \Delta y \to 0} \frac{F_y(x + \Delta x, y) \Delta y - F_y(x, y) \Delta y + F_x(x, y) \Delta x - F_x(x, y + \Delta y) \Delta x}{\Delta x \Delta y}
\]

\[
= \lim_{\Delta x, \Delta y \to 0} \left[ \frac{F_y(x + \Delta x, y) \Delta y - F_y(x, y) \Delta y}{\Delta x \Delta y} + \frac{F_x(x, y) \Delta x - F_x(x, y + \Delta y) \Delta x}{\Delta x \Delta y} \right]
\]

\[
= \lim_{\Delta x \to 0} \left[ \frac{F_y(x + \Delta x, y) - F_y(x, y)}{\Delta x} \right] + \lim_{\Delta y \to 0} \left[ \frac{F_x(x, y) - F_x(x, y + \Delta y)}{\Delta y} \right]
\]

\[
= \frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}.
\]

Since this expression is quite a bit more complicated than the one for divergence, here’s an intuitive explanation. The circle we are attempting to spin in the counterclockwise direction is shown in figure 6.11.
Each of the arrows along the side represents a component of the field $\mathbf{F}$ that can spin the circle. Clearly, if we want to spin the circle counterclockwise, the bottom arrow should be stronger than the top arrow and the right arrow should be stronger than the left arrow. This corresponds to the $x$ component of $\mathbf{F}$ decreasing as you move upwards and the $y$ component of $\mathbf{F}$ increasing as you move to the right. In other words, we want $\partial F_x / \partial y$ to be negative and $\partial F_y / \partial x$ to be positive. This is exactly the condition we need in order for the formula for the $k$-component of the curl to work out positive (which means the circle is spinning counterclockwise when viewed from above, the direction in which $k$ points).

You can follow essentially the same proof with rectangles oriented normal to the $i$ and $j$ vectors to find the $i$- and $j$-components of $\text{curl} \mathbf{F}$. (This is a little tedious. The most important part is the $k$-component, which gives you the curl in two dimensions. Both the $i$ and $j$ components involve $z$, which does not exist for a two-dimensional function.) Here is the full formula for the curl:

$$\text{curl} \mathbf{F} = \left( \frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} \right) \mathbf{i} + \left( \frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x} \right) \mathbf{j} + \left( \frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right) \mathbf{k}.$$ 

Since the derivations of the other two components are nearly identical to the derivation for the $k$-component, the curl is actually quite symmetric. In particular, consider the sequence $(x, y, z, x, y, z, \ldots)$. Find the letter corresponding to the component you want ($x = i$, $y = j$, $z = k$) in this sequence. Let the previous letter in the sequence be $a$ and the letter before that be $b$. Then the relevant component of the curl is

$$\frac{\partial F_a}{\partial b} - \frac{\partial F_b}{\partial a}.$$ 

Nevertheless, this is quite an ugly formula. But it just so turns out that this formula is also given by the expression

$$\text{curl} \mathbf{F} = \nabla \times \mathbf{F} = \begin{vmatrix} i & j & k \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_x & F_y & F_z \end{vmatrix}.$$ 

It’s amazing how pretty a bit of notation can make things! Go ahead and work out the cross product – you should get the formula given above.

### 6.6 Green’s Theorem
Prerequisite Knowledge. Understand the divergence theorem (see section 6.4) and curl (see section 6.5).

Green’s theorem is so analogous to the divergence theorem in two dimensions that its proof hardly needs any introduction. By the way, just as \( \text{div} \mathbf{F} \) must be continuous for the divergence theorem to apply, \( \text{curl} \mathbf{F} \) must be continuous for Green’s theorem to apply. A particular function where this restriction ends up being important is discussed in section 6.8.5.2.

<table>
<thead>
<tr>
<th>Divergence theorem in two dimensions</th>
<th>Green’s theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td>We want to prove ( \iint_R \text{div} \mathbf{F} , dA = \int_C \mathbf{F} \cdot ds ).</td>
<td></td>
</tr>
<tr>
<td>Remember that ( ds ) is a <strong>normal</strong> vector.</td>
<td></td>
</tr>
<tr>
<td>When we divide a region into two subregions, the net flux out of the region is equal to the sum of the net fluxes out of each of the two subregions, because any flux along the boundary line will be equal and opposite for the two subregions.</td>
<td></td>
</tr>
<tr>
<td>The definition of divergence is “net flux out divided by area, for very small area”. This could be written as ( \text{div} \mathbf{F} = \frac{d(\text{net flux out})}{dA} ).</td>
<td></td>
</tr>
<tr>
<td>To find the total flux out of a large number of these small regions, we must multiply each of them by ( dA ) and then add them together. In fact, this is exactly what the divergence theorem states: ( \iint_R \text{div} \mathbf{F} , dA = \int_C \mathbf{F} \cdot ds ).</td>
<td></td>
</tr>
<tr>
<td>( d(\text{net flux out}) ) \hspace{1cm} \text{total flux out}</td>
<td></td>
</tr>
<tr>
<td>( \text{curl} \mathbf{F} \cdot \mathbf{k} = \frac{d(\text{circulation})}{dA} ).</td>
<td></td>
</tr>
<tr>
<td>To find the total circulation around a large number of these small regions, we must multiply each of them by ( dA ) and then add them together. In fact, this is exactly what Green’s theorem states: ( \iint_R \text{curl} \mathbf{F} \cdot \mathbf{k} , dA = \int_C \mathbf{F} \cdot dx ).</td>
<td></td>
</tr>
<tr>
<td>( d(\text{circulation}) ) \hspace{1cm} \text{circulation}</td>
<td></td>
</tr>
</tbody>
</table>

Let’s clear up some technical details. Why did we have to choose the \( \mathbf{k} \)-component of the curl? We defined the curl using a limit definition analogous to the limit definition for divergence. However, this limit definition only gave one component of the curl. Which one? The component must be normal to the plane of the contour and oriented such that looking from the direction in which it is
pointing will make the line integral counterclockwise. The $\mathbf{k}$ unit vector is normal to the $xy$-plane that contains the contour, and it points out of the paper. So as long as we take the line integral in the counterclockwise direction, we can just substitute $\text{div} \, \mathbf{F}$ with $\text{curl} \, \mathbf{F} \cdot \mathbf{k}$ when proving Green’s theorem by analogy.

### 6.7 Stokes’ Theorem

**Prerequisite Knowledge.** Understand Green’s theorem (see section 6.6) and surface integrals.

Stokes’ theorem is a straightforward generalization of Green’s theorem to arbitrary surfaces in space rather than just flat regions in the plane, just as the gradient theorem is a generalization of the fundamental theorem of calculus to arbitrary curved paths rather than just straight lines.

In Green’s theorem, our circulation was within the $xy$-plane, so we needed the $\mathbf{k}$-component of the curl because it was normal to the $xy$-plane. For an arbitrary surface, the principle remains the same except that we need the component of the curl normal to the surface. Thus, the equation will change from

$$
\int_R \int \text{curl} \, \mathbf{F} \cdot \mathbf{k} \, dA = \int_C \mathbf{F} \cdot d\mathbf{x}
$$

to

$$
\int_R \int \text{curl} \, \mathbf{F} \cdot \mathbf{n} \, dS = \int_R \int \text{curl} \, \mathbf{F} \cdot dS = \int_C \mathbf{F} \cdot d\mathbf{x},
$$

where $R$ can now be an arbitrary region in space (well, sort of – no spheres [which don’t have edges] or Möbius strips [which don’t have proper unit normal vectors $\mathbf{n}$] please!) bounded by the closed contour $C$.

Note that the orientation of the line integral, as always, matters. In particular, it should be oriented counterclockwise when you are looking from the direction in which the normal vectors $d\mathbf{S}$ to the surface are pointing. Also, $\text{curl} \, \mathbf{F}$ should be continuous for Stokes’ theorem, just like for Green’s theorem.

#### 6.7.1 The Divergence of the Curl

**Prerequisite Knowledge.** Understand Stokes’ theorem (see section 6.7) and the divergence theorem (see section 6.4).

If you compute the expression $\text{div} \, \text{curl} \, \mathbf{F}$ and use the equality of mixed partials, you will see that it always equals $0$. (Try it!) That is, the curl has no divergence. But computing partial derivatives is really the most unintuitive way possible to prove this fact, especially given that the equality of mixed partials has no easily accessible graphical interpretation that I have been able to find.

However, Stokes’ theorem and the divergence theorem can be used together in a creative way to prove the same thing. We will start with a justification for one of the steps in the proof. Suppose that you know that $f(x)$ is a continuous function and also that $\int_a^b f(x) \, dx = 0$ for every possible $a$ and $b$. You can then conclude that $f(x) = 0$ always. (Convince yourself of this.) Similarly, if you know that $f(x, y, z)$ is a continuous function and also that $\int\int\int_Q f(x, y, z) \, dV = 0$ for every possible region $Q$, you can then conclude that $f(x, y, z) = 0$ always.

Now for the actual proof. Consider the scalar field $f(x, y, z) = \text{div} \, \text{curl} \, \mathbf{F}(x, y, z)$, where $\mathbf{F}$ is any vector field. Suppose we are given an arbitrary region in space $Q$. Then according to the divergence
theorem in three dimensions,
\[ \iiint_Q \text{div} \mathbf{F} \, dV = \iint_S \text{curl} \mathbf{F} \cdot d\mathbf{S}, \]
where \( S \) is the closed surface that bounds \( Q \). At this point, it will be useful to picture a region in space that you can easily divide into a “top half” and “bottom half”, such as a sphere. Draw an arbitrary contour dividing the surface into these two halves. (Of course, it does not matter what contour you pick. But it will be easier to keep track of which vectors are going which direction if both you and I are talking about the same halves of your surface.) The surface integral can then be divided into two:
\[ \iint_S \text{curl} \mathbf{F} \cdot d\mathbf{S} = \iint_{S_{\text{top}}} \text{curl} \mathbf{F} \cdot d\mathbf{S} + \iint_{S_{\text{bottom}}} \text{curl} \mathbf{F} \cdot d\mathbf{S}. \]

We can now apply Stokes’ theorem to both of these now non-closed surfaces. For the top half:
\[ \iint_{S_{\text{top}}} \text{curl} \mathbf{F} \cdot d\mathbf{S} = \int_C \text{curl} \mathbf{F} \cdot d\mathbf{x}, \]
where \( C \) is oriented counterclockwise because the normal vector is pointing out and up. Since both halves of the surface are bounded by the same contour, both surface integrals are equal to the same line integral – except that the orientation of the curve is the opposite! When you look from below, the normal vector is again outwards and pointing toward you, so the line integral is oriented counterclockwise. But since you are looking from below, the line integral is actually in the opposite direction to the original one. Since reversing the direction of a line integral reverses its sign, we have
\[ \iint_Q \text{div} \mathbf{F} \, dV = \iint_S \text{curl} \mathbf{F} \cdot d\mathbf{S} \]
\[ = \iint_{S_{\text{top}}} \text{curl} \mathbf{F} \cdot d\mathbf{S} + \iint_{S_{\text{bottom}}} \text{curl} \mathbf{F} \cdot d\mathbf{S} \]
\[ = \int_C \text{curl} \mathbf{F} \cdot d\mathbf{x} - \int_C \text{curl} \mathbf{F} \cdot d\mathbf{x} \]
\[ = 0. \]

According to our earlier discussion, since the integral of \( f(x, y, z) = \text{div} \mathbf{F}(x, y, z) \) is 0 over any region whatsoever, we must have \( \text{div} \mathbf{F} = 0 \) everywhere, which was to be demonstrated.

### 6.8 Conservative Vector Fields

**Prerequisite Knowledge.** Understand the gradient theorem (see section 6.2) and Stokes’ theorem (see section 6.7).

Vector fields (in both two and three dimensions) can have many properties – for instance, being divergence free, curl free, differentiable, and so on. One of the most important properties of a vector field is whether it is **conservative**. There are quite a few properties associated with conservative vector fields, and in this section we explore each of those properties.
6.8.1 Path Independence

The definition of conservative is path independent. These two terms are completely equivalent. But of course now we need to define what path independent means. If a vector field is path independent, then for any two points \( a \) and \( b \), any line integral between those two points will have the same value. For instance, the vector field \( \mathbf{F} = x \mathbf{i} + y \mathbf{j} \) is path independent. This means that the line integrals along the four paths pictured in figure 6.12 all have the same value.

![Figure 6.12: Four paths from \( \mathbf{a} = (0,0) \) to \( \mathbf{b} = (1,1) \) in the vector field \( \mathbf{F} = x \mathbf{i} + y \mathbf{j} \), all of which have the same value: 1. The fact that all four of these integrals are the same does not mean that \( \mathbf{F} \) is conservative, but it would be enough to make a reasonable hypothesis. You only need to find one integral with a different value to show that a vector field is not conservative, however.](image)

You should try computing these integrals; you will find that

\[
\int_C (x \mathbf{i} + y \mathbf{j}) \cdot d\mathbf{x} = \int_0^1 2t \, dt
\]

\[
= \int_0^1 2t^3 + t \, dt
\]

\[
= \int_0^1 \left[ \frac{3\pi}{2} \sin \left( \frac{3\pi}{2} t \right) \cos \left( \frac{3\pi}{2} t \right) + t \right] \, dt
\]

\[
= \int_0^{1/2} 5t \, dt + \int_1^{1/2} t \, dt
\]

\[
= 1.
\]

Of course, just because four line integrals between two points have the same value does not mean that every line integral will be the same for any given pair of two points. Later, we will see how to easily determine whether a vector field is path independent (conservative) without computing an uncountably infinite number of line integrals.

It is, however, much easier to show that a vector field is path dependent. You just have to find two paths between the same two points whose line integrals are different. For instance, for the field \( \mathbf{F} = -y \mathbf{i} + x \mathbf{j} \), the line integrals corresponding to the two paths shown in figure 6.13 have different values:

\[
\int_{C_1} (-y \mathbf{i} + x \mathbf{j}) \cdot (dx \mathbf{i} + dy \mathbf{j}) = \int_0^\pi \cos^2 \theta + \sin^2 \theta \, d\theta = \pi
\]
Try finding these integrals yourself. Your results should verify that the vector field \( \mathbf{F} \) is nonconservative.

\[
\int_{C_2} (-y \mathbf{i} + x \mathbf{j}) \cdot (dx \mathbf{i} + dy \mathbf{j}) = \int_0^\pi -\cos^2 \theta - \sin^2 \theta \, d\theta = -\pi
\]

Figure 6.13: Two paths between \( \mathbf{a} = (1, 0) \) and \( \mathbf{b} = (-1, 0) \) in the vector field \( \mathbf{F} = -y \mathbf{i} + x \mathbf{j} \) with different values. Of course, there may be other paths between \( \mathbf{a} \) and \( \mathbf{b} \) with the same values as one of these paths, but this is irrelevant: every path must give the same value in order for \( \mathbf{F} \) to be considered conservative.

6.8.2 Potential Functions

The potential function of a vector field \( \mathbf{F} \) is a scalar field \( f \) such that \( \text{grad} \, f = \mathbf{F} \). It is the multivariable version of an antiderivative (a function \( G \) such that \( G' = g \)). Some vector fields have potential functions and some do not. For instance, \( \mathbf{F} = x \mathbf{i} + y \mathbf{j} \) has the potential function \( f = \frac{1}{2}x^2 + \frac{1}{2}y^2 \) (check this), while \( \mathbf{F} = -y \mathbf{i} + x \mathbf{j} \) does not have any potential function. In fact, a vector field has a potential function if and only if it is conservative (path independent). We can prove this by first showing that if a vector field has a potential function then it must be path independent, and then showing that if a vector field is path independent then it must have a potential function.

The first part of the proof is easy. If \( \mathbf{F} \) has a potential function (that is, if \( \mathbf{F} = \text{grad} \, f \)) then according to the gradient theorem,

\[
\int_a^b \mathbf{F} \cdot d\mathbf{x} = \int_a^b \text{grad} \, f \cdot d\mathbf{x} = f(b) - f(a) = \text{constant}
\]

for any path between \( \mathbf{a} \) and \( \mathbf{b} \). Therefore, \( \mathbf{F} \) is path independent.

The second part is slightly more nuanced. Given that \( \mathbf{F} \) is path independent, we must find a potential function \( f \) such that \( \text{grad} \, f = \mathbf{F} \). In one dimension, the function

\[
G(x) = \int_a^x g(t) \, dt
\]
is an antiderivative of \( g(x) \) for any \( a \) according to the second fundamental theorem of calculus. Luckily, we can extend this almost directly to multiple dimensions: the function

\[
f(x) = \int_a^x F(z) \cdot dz
\]

(\( z \) has nothing to do with the \( z \)-axis; it is just a convenient letter) is a potential function of \( F \) for any \( a \) (assuming that \( F \) is path independent, because if it were not we would have to explicitly give a path for the integral). In any case, we can show that the gradient of this function is \( F(x) \) by considering two different paths from \( a \) to \( x \). (To have a comprehensible diagram, we’re assuming a two-dimensional space. In a three-dimensional space, you would have to consider three different paths.) These paths are shown in figure 6.14. (We’ll say that \( a = a_x i + a_y j \) and \( x = x i + y j \).)

Figure 6.14: Two equivalent ways of computing the potential function \( f \) we have constructed. Of course, since \( F \) is path independent, you can take any path from \( a \) to \( x \) and get the same value, but these two particular paths make it much easier to find the partial derivatives of \( f \) with respect to \( x \) and \( y \).

Because \( F \) is path independent, both of the line integrals pictured in figure 6.14 must have the same value and must also therefore have the same partial derivatives in both directions. So, we can find \( \partial f / \partial x \) from the left-hand diagram and \( \partial f / \partial y \) from the right-hand diagram.

For the left-hand contour,

\[
f(x) = \int_{a_y}^y F_y(a_x, z) \, dz + \int_{a_x}^x F_x(z, y) \, dz
\]

and

\[
\frac{\partial f(x)}{\partial x} = \frac{\partial }{\partial x} \int_{a_y}^y F_y(a_x, z) \, dz + \frac{\partial }{\partial x} \int_{a_x}^x F_x(z, y) \, dz = 0 + F_x(x, y) = F_x
\]

by recognizing that the first half of the integral does not depend on \( x \) and using the second fundamental theorem of calculus on the second half of the integral. For the right-hand contour,

\[
f(x) = \int_{a_x}^x F_x(z, a_y) \, dz + \int_{a_y}^y F_y(x, z) \, dz
\]

and

\[
\frac{\partial f(x)}{\partial y} = \frac{\partial }{\partial y} \int_{a_x}^x F_x(z, a_y) \, dz + \frac{\partial }{\partial y} \int_{a_y}^y F_y(x, z) \, dz = 0 + F_y(x, y) = F_y
\]
by recognizing that the first half of the integral does not depend on \( y \) and using the second fundamental theorem of calculus on the second half of the integral. So,

\[
\text{grad } f = \frac{\partial f}{\partial x} \mathbf{i} + \frac{\partial f}{\partial y} \mathbf{j} = F_x \mathbf{i} + F_y \mathbf{j} = \mathbf{F}.
\]

This proves that if a vector field is path-independent (conservative), then it must have a potential function. Since we already proved the converse, these two conditions are entirely equivalent.

### 6.8.3 The Conservation of Energy

At this point, we should discuss some nomenclature. Why are path-independent vector fields called “conservative”? Why is the anti-gradient called a “potential function”?

The answers to both of these questions lie in physics – and more specifically, in energy. What is energy? You will find a wide range of unsatisfying and sneakily worded answers from various sources, but really the answer is quite simple: fundamentally, energy is movement. The only “real” type of energy is kinetic energy. Specifically, an object with a mass \( m \) moving at a speed of \( v \) is defined to have a kinetic energy of

\[
K = \frac{1}{2}mv^2.
\]

Why this definition? It turns out that this particular definition is related very closely to another definition: that of work. When a force \( \mathbf{F} \) acts on an object moving along the path \( C \), the work done by that force on the object is defined to be

\[
W = \int_C \mathbf{F} \cdot d\mathbf{x}.
\]

We can prove what is called the work-kinetic energy theorem:

\[
W = \Delta K.
\]

Over any given time interval, the change in kinetic energy of an object with one force acting on it is equal to the work done by that force. The proof is just some simple but rather sneaky manipulations of the variables in the definition of work:

\[
W = \int_a^b \mathbf{F} \cdot d\mathbf{x} = \int_{t_a}^{t_b} m \mathbf{a} \cdot \frac{d\mathbf{x}}{dt} dt
\]

\[
= \int_{t_a}^{t_b} m \frac{d\mathbf{v}}{dt} \cdot \mathbf{v} dt
\]

\[
= \int_{t_a}^{t_b} \frac{1}{2} m \left( \frac{d\mathbf{v}}{dt} \cdot \mathbf{v} + \mathbf{v} \cdot \frac{d\mathbf{v}}{dt} \right) dt
\]

\[
= \frac{1}{2} m \int_{t_a}^{t_b} \frac{d}{dt} (\mathbf{v} \cdot \mathbf{v}) dt
\]

\[
= \frac{1}{2} m \int_{t_a}^{t_b} \frac{d}{dt} v^2 dt
\]

\[
= \frac{1}{2} m \left[ v^2 \right]_{t_a}^{t_b}
\]

\[
= \frac{1}{2} mv_b^2 - \frac{1}{2} mv_a^2
\]
\[ K_b - K_n = \Delta K. \]

At various points, we used a change of variables from position to time, Newton’s second law \( \mathbf{F} = m\mathbf{a} \), the fact that acceleration is the time derivative of velocity, the product rule for a dot product, the property that a vector dotted with itself gives the magnitude squared, and the fact every function is an antiderivative of its derivative.

The change in kinetic energy of an object is equal to the work done on it by the force. So what? Well, if the force field is path independent, the change in kinetic energy between two points is always the same, no matter what path is taken. This means that if the object has a given kinetic energy at one particular point (call it \( a \)), then its kinetic energy at any other point can only ever have one value for a given point. In particular, the kinetic energy is given by the potential function \( f(x) \) we wrote an expression for in section 6.8.2 (assuming that \( K = 0 \) at the starting point for the line integral). Thus, kinetic energy is a function of position for conservative force fields.

Again, so what? Well, if the kinetic energy is a function of position, then we can define another function of position, potential energy:

\[ \text{potential energy} = \text{total energy} - \text{kinetic energy}, \]

or

\[ U = E - K \]

in the commonly used nomenclature, where \( E \) can be any number we choose. (It doesn’t matter. Really.) But since \( E \) is a constant and \( K \) is a function of position, so is \( U \). This means if we say that every object has two types of energy – kinetic (a function of velocity) and potential (a function of position) – then the total amount of energy is constant. That is, energy is conserved. However, for path dependent (nonconservative) force fields, kinetic energy will not be a function of position and so we cannot define potential energy as a function of position. In that case, energy is not conserved. In other words, energy is conserved only in conservative fields. Furthermore, the potential energy is almost given by the potential function – the only difference is a negative sign and an added constant (the total energy). What coincidental nomenclature!

### 6.8.4 No Circulation

So far, we have shown that a vector field being conservative is equivalent to it being path independent, having a potential function, and obeying the conservation of energy. Yet another equivalent property is that conservative vector fields are circulation free. What does this mean? It is quite simple: we already mentioned in our definition of curl (section 6.5.2) that the circulation of \( \mathbf{F} \) around the closed contour \( C \) is equal to the line integral

\[ \int_C \mathbf{F} \cdot d\mathbf{x}. \]

Take a look at figure 6.15.

We have picked two arbitrary points, \( a \) and \( b \), and two arbitrary paths between them, \( C_1 \) and \( C_2 \). (We denote the respective paths in the opposite directions as \( C'_1 \) and \( C'_2 \).) In order to show that path independence and lack of circulation are equivalent conditions, we will show first that if a vector field is path independent, then it must have no circulation; then that if a vector field has no circulation, then it must be path independent.
Figure 6.15: A closed contour containing \( \mathbf{a} \) and \( \mathbf{b} \), which can also be viewed as two distinct paths between \( \mathbf{a} \) and \( \mathbf{b} \). There are a total of four paths shown between \( \mathbf{a} \) and \( \mathbf{b} \): two from \( \mathbf{a} \) to \( \mathbf{b} \) (\( C_1 \) and \( C_2 \)), and two from \( \mathbf{b} \) to \( \mathbf{a} \) (\( C_1' \) and \( C_2' \)).

For the first condition, we know from path independence that

\[
\int_{C_1} \mathbf{F} \cdot d\mathbf{x} = \int_{C_2} \mathbf{F} \cdot d\mathbf{x}
\]

However, because reversing a line integral reverses its sign, we know that

\[
\int_{C_2} \mathbf{F} \cdot d\mathbf{x} = -\int_{C_2'} \mathbf{F} \cdot d\mathbf{x}
\]

Therefore, the circulation along the entire contour, going counterclockwise, is

\[
\int_{C_1} \mathbf{F} \cdot d\mathbf{x} + \int_{C_2} \mathbf{F} \cdot d\mathbf{x} = \int_{C_1} \mathbf{F} \cdot d\mathbf{x} - \int_{C_1'} \mathbf{F} \cdot d\mathbf{x} = 0.
\]

Because this result would hold for an arbitrary closed path divided into two halves, \( \mathbf{F} \) must be circulation free.

For the second condition, we know from lack of circulation that

\[
\int_{C_1} \mathbf{F} \cdot d\mathbf{x} + \int_{C_2} \mathbf{F} \cdot d\mathbf{x} = 0.
\]

Since \( \int_{C_2} \mathbf{F} \cdot d\mathbf{x} = -\int_{C_2'} \mathbf{F} \cdot d\mathbf{x} \), we have that \( \int_{C_1} \mathbf{F} \cdot d\mathbf{x} - \int_{C_2} \mathbf{F} \cdot d\mathbf{x} = 0 \), or \( \int_{C_1} \mathbf{F} \cdot d\mathbf{x} = \int_{C_2} \mathbf{F} \cdot d\mathbf{x} \).

Therefore, a vector field has no circulation if and only if it is conservative.

6.8.5 No Curl

6.8.5.1 The Property

Recall that curl and circulation are closely related by Green’s theorem. In particular,

\[
\iint_R \text{curl} \mathbf{F} \cdot \mathbf{k} \, dA = \oint_C \mathbf{F} \cdot d\mathbf{x}.
\]

You can see from this equation that if \( \text{curl} \mathbf{F} \) is zero everywhere within a region \( R \), then the circulation around \( R \) must be zero. So if you have a region \( R \) inside which \( \text{curl} \mathbf{F} = 0 \), then the circulation around any closed path inside \( R \) must be zero. Therefore, \( \mathbf{F} \) has no circulation and must be conservative!

Furthermore, if \( \mathbf{F} \) is conservative – that is, if the circulation around every closed path inside \( R \) is zero – then the integral of \( \text{curl} \mathbf{F} \) over every region inside \( R \) must be zero, and it follows [by reasoning analogous to that which we used in our proof that the curl has no divergence (section 6.7.1)] that \( \text{curl} \mathbf{F} = 0 \) everywhere inside \( R \).

In this proof, we used Green’s theorem, which is of course only valid in two dimensions. In three dimensions, however, it is quite simple to substitute Stokes’ theorem, and the same result follows.

6.8.5.2 The Caveat

Great! So curl free and circulation free are equivalent, just like everything else? Not quite. What if \( \mathbf{F} \) isn’t defined at a certain point? In that case, no circulation still implies no curl. If there is
no circulation around any contour \( C \), then the integral of the curl over any region \( R \) is zero. This implies that the curl is zero. But on the other hand, no curl does not necessarily mean no circulation. To prove that there is no circulation, we have to consider every possible contour \( C \), including the ones circling the point at which \( \mathbf{F} \) is not defined. And we cannot apply Green’s theorem to a region unless \( \text{curl} \mathbf{F} \) is continuous everywhere inside \( R \).

You might think that we could simply choose a donut-shaped region \( R \) so that the problem point is inside the hole, and therefore not a part of \( R \). But again, we have to consider all possible paths within \( R \), including the ones that circle the hole. If a contour circles the hole in \( R \), then the region it surrounds will not be entirely within \( R \). We are back to the same problem, because we cannot apply Green’s theorem to every possible contour. (By the way, everything else we’ve discussed about conservative vector fields holds true in a region with a hole – just not using Green’s theorem to prove a lack of circulation.)

All of this may have seemed a little theoretical. Here is a practical example for you to work. Take the vector field

\[
\mathbf{F} = \frac{y}{x^2 + y^2} \hat{i} - \frac{x}{x^2 + y^2} \hat{j}.
\]

Find \( \text{curl} \mathbf{F} \). You should see that it is zero everywhere. Then compute the line integrals along the two paths shown in figure 6.13 on page 61. You should find that they are different. You can also find the line integral around the unit circle – a closed path – which should be nonzero. These properties – path dependence and circulation – would be contradictions of the fact that vector fields without curl are conservative, if it were not for the fact that \( \mathbf{F} \) is not defined (with a nonremovable discontinuity) at the origin. If you wanted to conclude that \( \mathbf{F} \) were conservative, you could not use curl – you would have to show that it has a potential function, that it is path independent, or that it lacks circulation. Of course, the latter two do not work, as you proved when computing the path integrals earlier. What about a potential function? Try to use integration to find one – you should find that

\[
f = \arctan \left( \frac{x}{y} \right).
\]

Isn’t this a contradiction? No, because \( f \) is not defined anywhere along the \( x \)-axis. Try computing the gradient of

\[
f = - \arctan \left( \frac{y}{x} \right).
\]

You should obtain the same value for \( \mathbf{F} \). This \( f \), of course, is not defined along the \( y \)-axis. You can find many potential functions for \( \mathbf{F} \), but you will never be able to find one that is defined all the way around the origin.

What is happening here, intuitively? There is clearly circulation, so why does that not lead to curl? Effectively, all of the curl has been compressed into a single point at the origin, which explains why any curve not containing the origin will have zero circulation, but any curve circling the origin will have a nonzero circulation. Instead of lots of curl spread over a region adding together to make circulation, as is typically the case when applying Green’s theorem (or Stokes’ theorem), there is an infinitely dense “bundle of curl” centered at the origin which adds a discrete amount of circulation to any path that contains it. And since infinity is beyond what a regular function can comprise, the vector field (or function) \( \mathbf{F} \) is not defined at the origin. Instead of being pulled upwards to infinity, like a function with an asymptote would be, the components of the vector field are pulled in different directions infinitely far – and so they cannot be consistently defined at the origin.
6.8.5.3 The Topology

In the previous section, we explored the fact that lack of curl does not necessarily imply lack of circulation when the region has a hole. Let us now specify more precisely which regions you can apply the theorem to. They are called simply connected regions. There are several equivalent definitions of a simply connected region (which can have two, three, or more dimensions):

- a region that is contiguous and does not have any “holes that go all the way through it”
- a region such that (1) any two points in the region can be connected by a curve entirely within the region, and (2) for any closed loop contained entirely within the region, the loop can be shrunk to a point continuously while staying entirely within the region
- a region such that (1) any two points in the region can be connected by a curve entirely within the region, and (2) any curve within the region can be continuously deformed into any other curve within the region while staying entirely within the region

Here are some examples:

- Any non-contiguous region is not simply connected.
- Any contiguous two-dimensional region without holes is simply connected.
- Any two-dimensional region with one or more holes is not simply connected.
- A circular band is not simply connected.
- A teacup with a handle is not simply connected.
- A solid sphere is simply connected.
- A solid sphere with a hole in the middle (like a rubber ball) is simply connected.
- A solid sphere with a hole drilled all the way through it is not simply connected.

Much of higher math is translating rather technical definitions into intuitive concepts you can visualize. It is not difficult to get an accurate mental picture of what it means to be simply connected, but it will take some good thought. Try it!

Knowing the definition of simply connected regions may not seem important, but it is actually very useful because you can only conclude that a curl-free vector field is conservative if the region is simply connected! (And, of course, it is usually much easier to check that a vector field is curl free than to check if it has a potential function – to say nothing of checking path independence or lack of circulation.)

6.8.6 Summary

For any region \( R \):

- \( \mathbf{F} \) is conservative \( \iff \) \( \mathbf{F} \) is the gradient of a scalar field
- \( \mathbf{F} \) is conservative \( \iff \) \( \mathbf{F} \) is path independent
- \( \mathbf{F} \) is conservative \( \iff \) \( \mathbf{F} \) has no circulation
• **F** is conservative and continuous $\implies$ **F** has no curl

• **F** has no curl and **R** is simply connected $\implies$ **F** is conservative over that region

An elementary consequence of these facts is that \textbf{curl grad }f = 0, because \textbf{grad }f is the gradient of the scalar field \( f \), so \textbf{grad }f is conservative and consequently has no curl.