One-variable Calculus

Toby Bartels
MATH-1600
2022 Spring

Welcome to Calculus! Here are my supplemental notes for one-variable Calculus, giving alternative ways to think about some things, practical advice, and sometimes more theoretical detail.

This does not cover everything that you need to know; you should also have the official course textbook, which is the 4th Edition of *University Calculus: Early Transcendentals* by Hass et al published by Addison–Wesley (Pearson). There are also some references in these notes to that textbook. Conversely, there is some material in here that you *don't* need to know, although I hope that it will be helpful; I'll generally make a note of that when it happens. Most of the stuff at the end won't be needed until Calculus 2.

Contents

Detailed contents

1 Preliminaries

Before beginning this class, you should be familiar with the basic algebraic properties of real numbers and real-valued functions of real numbers.

1.1 *Numbers*

By default, all of the numbers that we work with will be real numbers. (Most of Calculus applies just as well to complex numbers, but a complete understanding of Calculus in even one complex variable requires some ideas from multivariable Calculus, which these notes do not cover.) In particular, if a is a negative number, then $\sqrt[n]{a}$ is undefined when n is an even integer and negative when n is an odd integer. More generally, if a is a negative number, then a^p is defined if and only if p is a rational number whose denominator in lowest terms is odd; in this case, a^p is positive if the numerator of p is even and negative if the numerator of p is also odd. Note that $(a^2)^{1/2} = \sqrt{a^2} = |a|$ in general, while $a^{2\cdot 1/2} = a^1 = a$, which is different from $|a|$ when a is negative, so the rule that $(a^x)^y = a^{xy}$ does *not* hold in general (although it *does* hold when a is a positive number).

Although 0^x is undefined whenever x is negative (because this amounts to dividing by zero), we need to define $0^0 = 1$ in order to make some formulas work correctly. Many textbooks say that 0^0 is undefined (although ours is a little coy about this), but the general theory of polynomials and related concepts is simpler when $0^0 = 1$. (That is what the textbook implicitly assumes in its definition of power series on page 543 at the beginning of Section 9.7. See also the discussion of power series starting on page 61 of these notes.) It's possible to take a more nuanced approach, where 0^x is 1 when x is an *integer*-valued variable with the value 0 while 0^x is undefined when x is a *real*-valued variable with the value 0; however, this makes the meaning of 0^0 ambiguous without context, so for simplicity, I will just say that $0^0 = 1$. Nevertheless, this will require some care when it comes to rules for evaluating limits; that's why 0^0 still appears among the indeterminate forms on page 246 at the beginning of Section 4.5 of the textbook and on page 16 of these notes.

When we use trigonometric operations, they will always apply to angle measures in radians. Actually, it's best to think of these as operations on pure numbers, with the geometric application to angles as just one use of them. So $\sin x$ and $\cos x$ are defined for any real number x, $\sin(x + 2\pi)$ is always the same as $\sin x$, etc. Also, for the inverse trigonometric functions, I write $\operatorname{asin} x$ for the unique real number such that $-\pi/2 \leq \operatorname{asin} x \leq \pi/2$ and $\sin(\operatorname{asin} x) = x$ (if there is any such number at all, which there will be if and only if $-1 \leq x \leq 1$); this number is also variously written $\arcsin x$ (as in Chapter 1 of the textbook), $\operatorname{Sin}^{-1} x$, or $\sin^{-1} x$ (as in Chapter 3 of the textbook). Similarly, $\operatorname{acos} x$ is the unique real number such that $0 \leq \operatorname{acos} x \leq \pi$ and $\cos(\operatorname{acos} x) = x$ (if there is any such number). We can similarly define atan, acot, asec, and acsc, although these don't get as much use. Note that I use $0 \leq \operatorname{asec} x \leq \pi$ and $-\pi/2 \leq \operatorname{acsc} x \leq \pi/2$; some Calculus textbooks do this differently, but I am agreeing with our textbook in this respect.

The main difference between my approach to Calculus and the textbook's is that I make more use of *differentials*. Calculus was originally developed using differentials, and many calculations are easier to do this way. Furthermore, differentials are often used in applications, especially (but not only) to physics. They fell out of fashion with mathematicians towards the end of the 19th century, when Calculus was first put on a rigorous logical foundation, because this foundation did not include differentials. However, a rigorous logical development of differentials as well had been achieved by the early 20th century, so there is no longer any reason to avoid them. You can do almost everything with the textbook's methods if you want, but I encourage you to try using differentials. (This will be especially fruitful if you go on to take Calculus 3, where differentials are even more convenient.)

A related (but distinct) issue is the question of infinitely small (but nonzero) numbers. We say that a number is infinitely small, or *infinitesimal*, if its absolute value is less than 1, less than 1/2, less than 1/3, etc. In the real number system as we now understand it, the only infinitely small number is 0; however, in the early days of Calculus, people reasoned in terms of nonzero infinitesimal numbers (and their reciprocals, which are infinitely large numbers) quite often. I will discuss this occasionally, because they can be useful for intuitive understanding, but this is entirely optional; I'll make no attempt at a complete or rigorous discussion of such numbers, although I'll make sure that everything that I say about them is at least true. (Infinitesimal numbers were the last concept to be made fully rigorous, but even so, this was done in 1960, probably well before any of us was born.)

1.2 *Sets*

Geometrically, a **set** of real numbers is a region within the number line; for each number c, you should be able to say (in principle) whether c is in the set or not. That is, if c is a number and S is a set, then this claim is a mathematical statement that may be true or false. When it is true, we write $c \in S$ and say that c is an *element* and *member* of S, that c *belongs* to S, and that S *owns* c. Although one can talk about sets whose elements are anything at all (even other sets) rather than just real numbers, the default meaning of 'set' in this class is a set of real numbers. Note that both the entire real line (written $\mathbf{R}$, $\mathbb{R}$, or $(-\infty, \infty)$), which owns every real number, and the empty set (written $\varnothing$ or $\{\}$), which owns nothing at all, count as extreme examples of sets.

In general, you can define a set by picking a variable (say x) to stand for an arbitrary real number and writing down a statement about that number (using that variable) so that x belongs to the set if and only if the statement is true. For example, you might define a set S by saying that, for each real number x, $x \in S$ if and only if $x < 2$. (Note that 'if and only if' goes both ways: if $x \in S$, then $x < 2$; and if $x < 2$, then $x \in S$.) You can write this as $S = \{x \mid x < 2\}$, or $S = \{x \in \mathbf{R} \mid x < 2\}$ to emphasize that it's a set of real numbers. Or if you don't want to give the set a name like S, then you can refer to the set directly as $\{x \mid x < 2\}$. Then given any real number c, you know that $c \in \{x \mid x < 2\}$ if and only if $c < 2$. For example, $1 \in \{x \mid x < 2\}$, because $1 < 2$; but $3 \notin \{x \mid x < 2\}$, because $3 \not< 2$ (where the slashes indicate that something is *not* true).

We will often have to deal with **intervals**, which are particular sets of real numbers, so there is a special notation for them. If a and b are real numbers with $a < b$, then $[a, b]$, $[a, b)$, $(a, b]$, and (a, b) are all sets (the intervals *from a to b*, or with a and b as *endpoints*), consisting of all of the numbers strictly between a and b, as well as possibly the endpoints a and b themselves; an endpoint belongs to the interval if the bracket on that side is square but not if it is round. We can also use $-\infty$ in place of a or ∞ in place of b (or both), to indicate that the interval continues forever in that direction; but because $-\infty$ and ∞ are not real numbers, the brackets next to them must always be round. In other words:

$$\begin{aligned} [a,b] &= \{x \mid a \le x \le b\}; & [a,b) &= \{x \mid a \le x < b\}; & [a,\infty) &= \{x \mid x \ge a\}; \\ (a,b] &= \{x \mid a < x \le b\}; & (a,b) &= \{x \mid a < x < b\}; & (a,\infty) &= \{x \mid x > a\}; \\ (-\infty,b] &= \{x \mid x \le b\}; & (-\infty,b) &= \{x \mid x < b\}; & (-\infty,\infty) &= \mathbf{R}. \end{aligned}$$

We call $[a, b]$, $[a, \infty)$, $(-\infty, b]$, and $(-\infty, \infty)$ **closed** intervals; they include all of the endpoints that they can. Conversely, we call (a, b), (a, ∞), $(-\infty, b)$, and $(-\infty, \infty)$ **open** intervals; they include none of their endpoints. (Notice that $[a, b)$ and $(a, b]$ are neither open nor closed, while $(-\infty, \infty)$ is both.) Also, the intervals that don't involve any kind of infinity are called **bounded** intervals. In particular, the closed bounded intervals of the form $[a, b]$ are called **compact** intervals. These will all be useful notions from time to time.

Although I said above that $a < b$ for the endpoints of an interval, we also allow $a = b$ for compact intervals; however, $[a, a]$ is more commonly written simply $\{a\}$; that is, curly braces with the single element a listed within them. (You can generalize this to a list of any finite number of elements, separated by commas, but then the set will no longer be an interval.) If you're talking about an interval from a to b and want to ensure that $a < b$, then you can speak of a **nontrivial interval**. This is usually just a technicality, however.

1.3 *Functions*

One difference between these notes and the textbook is that I will never be sloppy with function notation.

In an expression such as

$$y = f(x),$$

the variables x and y stand for real numbers, while the variable f stands for a function. (Usually this variable is actually a constant, because f always refers to the same function throughout the problem; if I want to emphasize that this is so, then I'll say that f is a *fixed* function.) A function is not a number but rather a process for turning one number into another. When speaking of specific numbers, this is usually not a problem; for example, $f(2) = 4$ means that the function f is a process that (among other things) turns the number 2 into the number 4.

The statement that $f(x) = x^2$ is more ambiguous; in a context where the variable x already appears, this means that the function f is a process that (among other things) turns the number x (whatever number that is) into the number x^2. But in a context where x does not already have a meaning, this statement usually means that the function f is a process that turns *every* real number into its square, which is a complete description of the function. In this case, it is better to say something like

$$f(x) = x^2 \text{ for all } x,$$

and I will usually say something like this.

Another way to completely describe this function is to write

$$f = (x \mapsto x^2).$$

This is analogous to defining a set S as $S = \{x \mid x < 2\}$; in each case, you introduce a new *dummy variable* and then you either give an expression (to define a function) or else you give an equation, inequality, or other statement (to define a set), in each case using that dummy variable. You can even do this without giving the function a name, by (for example) referring to the function $(x \mapsto x^2)$ (just like referring to the set $\{x \mid x < 2\}$); this is called **anonymous function** notation. Although the textbook does this with sets, it never does this with functions, so I won't do it much either. It can be handy, however.

The real problem is when the same symbol is used both to refer to a function and to its output value, as in

$$A = A(x),$$

which you might see (for example) in a problem in which the ***a***rea of some shape depends on something else. I will never do this! Either I will use A to refer to the area itself, or I will use A to refer to the function that indicates how this area depends on x (whatever that may be in the situation), but I will not use the same symbol for both of these. If I want to refer to both of them, then I will use two different symbols; most of the time, however, it's enough to have a symbol for the area itself and to leave the function unnamed. (The *evaluation notation* described on page 6 can help with this.)

When we cover derivatives later on (in Chapter 3), you will learn various symbols used for this concept; and when $y = f(x)$, then I will also write

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f'(x).$$

(What this means is explained on page 22 these notes.) The textbook will sometimes write y' or $\mathrm{d}f/\mathrm{d}x$ in this situation, but I never will, and this is important to ensure that the ordinary rules of algebra continue to apply to such expressions. (For example, you can multiply both sides of the equation above by $\mathrm{d}x$ to get $\mathrm{d}y = f'(x)\,\mathrm{d}x$, which would be difficult to do correctly using the wrong symbols.) I will not count it against you if you are as sloppy as the textbook about this, because I don't think that it's fair to require you to do more than the textbook writers do; however, if you get confused by your notation and make a mistake, then that will count against you! So I encourage you to follow my lead and use precise notation.

1.4 *Variables*

In Calculus, we study *variable* quantities, that is quantities whose values may vary (or change).

In Algebra, we often use the word 'variable' to refer to any quantity whose value we don't know, even if this value is fixed and never changes throughout the problem. In fact, the standard Algebra problem, solving an equation such as $2x + 3 = 5$, involves figuring out the value of the variable; so it had only one value all along, and we just had to figure out what it was. So if x is a variable in an Algebra problem, and at some point we decide that the value of x is 1, then this may well mean that x is 1 throughout the entire problem. (That's not always the case in Algebra, but it often is.)

In Calculus, we take the word 'variable' more seriously. If x is a variable in a Calculus problem, then x might be 1 at some point, but it may well be 6 at some other point in the problem. (And more often than not, it will take all of the values in between 1 and 6 along the way, such as $1\frac{1}{2}$, π, and 5.789.) Furthermore, if x and y are two variables that appear in the same problem, then the value of y will usually change as the value of x changes. Calculus is primarily about exactly this sort of thing: *how* one quantity changes as another quantity changes.

In the simplest cases, it turns out that y is a function of x; that is, there is a fixed function f such that $y = f(x)$ remains true as x and y vary. Calculus textbooks generally try to fit everything into this mould, but it doesn't always come out like this naturally. Often, you know that both x and y are changing, but it's not obvious that the value of x at some point is enough information to figure out the value of y at that point; yet when you write $y = f(x)$, you're assuming that it is enough information.

Nearly all of the time, however, we can assume that there is some variable t, called the *independent variable*, such that every other variable in the problem is a function of t. That is, if x and y appear in the problem, then there are fixed functions g and h such that $x = g(t)$ and $y = h(t)$ throughout the problem. (Then x and y are called *dependent variables*, since their values depend on the values of t, through the functions g and h.) If it also happens that $y = f(x)$ throughout the problem, then this means that h is the composite function $f \circ g$; but if that doesn't happen, then at least we still have g and h.

However, this variable t might not show up directly! Calculus books will usually tell you (especially in word problems) that it's necessary to pick an independent variable from among the variables that appear in the problem, but really it's enough to informally visualize the range of variation in the problem, and you can treat all of the variables on an equal footing. All the same, for the sake of formal definitions, I will assume that there is an independent variable t and that every other variable is a function of it, even though in practice we don't have to identify it. (Of course, you don't have to call the independent variable 't', but I usually will, just to have a consistent name.)

If we're not going to refer directly to t, then we're not going to refer directly to g and h either, only to the quantities x and y; so we need some way to refer to the values of these quantities without referring to the functions that determine them. Here is how we do it formally:

$$\text{If } u = f(t), \text{ then } u|_{t=c} = f(c).$$

(This is called **evaluation** notation.) More generally, if P is some statement that is only true once, then P implies the statement $t = c$ for some value of c, so we can make sense of $u|_P$. Even if P is a statement that might not only be true once, as long as every possible value of $u|_P$ is the same, then we can still make sense of $u|_P$. Finally, even if there are different possible values of $u|_P$, then the value of $u|_P$ still varies, but at least it doesn't vary as much as u itself, since there are now fewer possibilities.

This all sounds very abstract (because it is), but the concrete application is straightforward; here are some examples:

$$\begin{aligned} x|_{x=5} &= 5, \\ (2x+3)|_{x=4} &= 2(4) + 3 = 11, \\ (2x+3y)|_{\substack{x=4,\\ y=5}} &= 2(4) + 3(5) = 23. \end{aligned}$$

Taking the last of these for example, there is no need to think about what t is when $x = 4$ and $y = 5$, and indeed without considering how x and y depend on this unspecified independent variable t, the value of t is impossible to know. Nevertheless, we know that no matter what t may be, if $x = 4$ and $y = 5$ at that value of t, then $u = 2x + 3y$ is definitely $2(4) + 3(5) = 23$ at that same value of t, and that is enough. So all that you have to do in practice is to plug in the given values and perform the given calculation.

Sometimes (generally only in the middle of a problem or in something theoretical) you can't work out the value completely; for example,

$$(2x + 3y)|_{x=4} = 2(4) + 3(y|_{x=4}) = 8 + 3y|_{x=4}.$$

If we don't know anything more about the relationship between x and y, then we don't know the value of y when $x = 4$, so this is all that we can say in this example, but at least we were able to work out part of it.

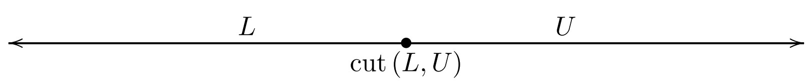
1.5 *Completeness of the real line*

In this course, we work with the real numbers, which are supposed to correspond to points on a number line. Ultimately, all of the properties of real numbers derive from intuitive geometric properties of points on a line. For example, the arithmetic operations of addition, subtraction, multiplication, and division can be defined in terms of changes of position and scale on the number line. The order relation between real numbers ($<$ and $>$) also derives from relative position on a line. (You have to specify the numerical values of at least two points, such as 0 and 1, in order to make a geometric line into a number line, but once you have those two points, then everything else follows.)

The most advanced of the fundamental properties of the number line is its ***completeness***. In this course, the topic of completeness is optional, so you can skip this section if you want; however, you'll need to understand it if you want to understand *why* everything in this course is true.

There are many ways to express completeness, but my favourite is this:

> If you pick out two nonempty regions of the number line, one on the left called L and one on the right called U, which don't overlap but otherwise cannot be extended further, then there is a single point between them, called $\operatorname{cut}(L, U)$, the ***cut*** between L and U.



We can make this logically precise (in terms of the order relation on real numbers): Suppose that L and U are two sets of real numbers (making precise what regions of the number line are), with these properties:

- There is some $r \in L$ and some $s \in U$ (which is what it means for L and U to be nonempty);
- If $r \in L$ and $s \in U$, then $r < s$ (which is what it means for L to be on the left and U on the right without overlapping);
- If $r < s$, then $r \in L$ or $s \in U$ (which is what it means to say that L and U cannot be extended further).

(Note that 'or' in math, as here, normally includes the possibility of both.) Then there exists a real number $\operatorname{cut}(L, U)$ with this property:

- If $r \in L$ and $s \in U$, then $r \leq \operatorname{cut}(L, U) \leq s$ (which is what it means for $\operatorname{cut}(L, U)$ to be between L and U).

A couple more important properties follow from what was said above:

- The number $\operatorname{cut}(L, U)$ is the *only* real number between L and U;
- If $r < \operatorname{cut}(L, U) < s$, then $r \in L$ and $s \in U$.

The point of all this is to be able to prove that a real number exists. For example, in order to prove rigorously that every real number c has a cube root $\sqrt[3]{c}$ (and has anybody ever showed you why this is true or did you just take it on faith?), you first define L as $\{x \mid x^3 < c\}$ and U as $\{x \mid x^3 > c\}$, check that L and U have the necessary properties listed above (which takes a bit of work with algebra), conclude that $\operatorname{cut}(L, U)$ exists with the properties listed above, and check (using those properties) that $\operatorname{cut}(L, U)^3 = c$ (which takes a lot more work with algebra). Thus, this cut is the cube root $\sqrt[3]{c}$.

This method of proving that a real number exists is also practical, because it shows us how to approximate its value as closely as we like. For example, to approximate $\sqrt[3]{2}$ to 4 decimal places, you look at some nearby possibilities, such as $1.0001, 1.0002, 1.0003, \ldots, 1.9997, 1.9998, 1.9999$. Somewhere in this list are two numbers right next to each other, one of which has a cube less than 2 (so it's in L) and one of which has a cube greater than 2 (so it's in U). Then we approximate $\sqrt[3]{2}$ to 4 decimal places by saying that it's in between these two numbers. (As it happens, these two numbers are 1.2599 and 1.2600; also, $1.25995^3 > 2$, so $\sqrt[3]{2}$ rounds down, to 1.2599.) There are more efficient ways to calculate cube roots (such as Newton's Method, described in Section 4.3), but this proof that they exist at least gives *one* way to calculate them, to start with.

I will only refer to this property of real numbers occasionally, when explaining why some number exists. The main point is that you know that a number exists if you can approximate it as closely as you like, the way that I approximated $\sqrt[3]{2}$ to 4 decimal places. Checking all of the detailed requirements is not usually really necessary to understand what's going on.

2 Limits and continuity

There are four main operations considered in Calculus: limits, derivatives (or differentials), integrals (or antidifferentials), and sums of infinite series. (The last of these is only covered in Calculus 2.) Here we will look at the first one: limits. These are also closely related to the concept of continuity, which is actually the easiest concept to define rigorously.

2.1 *Continuity*

In Calculus, we not only study variable quantities; we study quantities that are *continuously* varying. This implies in particular that a quantity y that varies from 1 to 6 will typically pass through $1\frac{1}{2}$, π, and 5.789, and everything else in between.

In real life, we can never measure or fix the value of a such a quantity y exactly, down to the last decimal place; after all, there are infinitely many decimal places, but we can only do a finite amount of work. So, it is key to the study of real numbers that we can *approximate* them to any finite number of decimal places, among other ways. (That is what the stuff about cuts in Section 1.5 accomplishes.)

Also in Calculus, we study how one quantity y varies along with another quantity x. The most straightforward way in which this can happen is when y is a *function* of x; if f is the function, then $y = f(x)$. But in practice, we only know x and y *approximately*, so if we only use an approximate value of x, then $f(x)$ should still be an approximate value of y. For example, suppose that $f(x) = x^2$ for all x; if you know that x is approximately 5, then you know that $y = f(x)$ is approximately $5^2 = 25$.

This doesn't work with every function! For example, suppose that g is the piecewise-defined function

$$g(x) = \begin{cases} x^2 & \text{for } x \leq 5, \\ x^3 & \text{for } x > 5; \end{cases}$$

if you only know that x is approximately 5, then you really don't know if $g(x)$ is approximately $5^2 = 25$ or approximately $5^3 = 125$. To be sure, if you know that x is *exactly* 5, then you know that $g(x)$ is $5^2 = 25$ (exactly); but it's no good if you only know x approximately.

In these examples, we say that g has a **discontinuity** at 5, while f is **continuous** at 5. (In fact, f is continuous everywhere, while g is continuous everywhere except at 5.) So the idea is this:

> A function f is **continuous** at a real number c if, whenever $x \approx c$ (meaning that x is approximately equal to c), $f(x) \approx f(c)$.

So if you only know that $x \approx c$, then that's enough information to know $f(x)$ approximately (specifically, that $f(x) \approx f(c)$).

Actually, we should take care about where f is defined. Sometimes Calculus textbooks say that f has a discontinuity at c if f is undefined at c (that is if $f(c)$ does not exist), and sometimes they don't; but in any case, f is not continuous there: f must be defined first in order to be continuous. On the other hand, if f is undefined at x, then we don't hold that against f; for example, we want to say that $f(x) = \sqrt{x}$ is continuous at 0, even though $f(x)$ does not exist (as a real number) when $x < 0$. So a more careful definition is this:

> A function f is **continuous** at a real number c if $f(c)$ exists and, whenever $x \approx c$ and $f(x)$ exists, $f(x) \approx f(c)$.

This is still not a completely rigorous definition, because it doesn't explain how close we need to be to say that one quantity is approximately equal to another. (Basically, the answer is this: as close as you need, and as close as you want.) But I will save that for the next section. Already, this basic idea should be enough to allow you to judge continuity of a function from its graph.

To judge continuity of a function from a formula, it's convenient to know that any function is continuous (wherever it is defined) if it has a formula that uses only these operations: addition, subtraction, multiplication, division, absolute values, opposites, reciprocals, raising to powers when the exponent is constant or the base is always positive, extracting roots when the index is constant or the radicand is always positive, logarithms, trigonometric functions, and inverse trigonometric functions. These are pretty much all of the functions that you ever deal with!

So, the exceptions in practice are much rarer: exponentiation where the exponent varies and the base can be zero or negative, roots where the index varies and the radicand can be zero or negative, and piecewise-defined functions. Of these, only piecewise-defined functions are likely to come up. These functions *can* be continuous, but only if the values agree on both sides whenever two pieces join. So for example, while

$$g(x) = \begin{cases} x+1 & \text{for } x < 2, \\ x+3 & \text{for } x \geq 2 \end{cases}$$

has a discontinuity at $x = 2$,

$$h(x) = \begin{cases} x+1 & \text{for } x < 2, \\ 5-x & \text{for } x \geq 2 \end{cases}$$

is continuous at $x = 2$ (and so everywhere), because $2 + 1 = 5 - 2$. The precise theorem is that, if f and g are functions that are each continuous at a number c, and if $f(c) = g(c)$, then the piecewise-defined function h defined by

$$h(x) = \begin{cases} f(x) & \text{for } x < c, \\ g(x) & \text{for } x \geq c \end{cases}$$

(or by $f(x)$ for $x \leq c$ and $g(x)$ for $x > c$, because this gives the same result), then h is also continuous at c.

2.2 *Defining continuity*

Returning to the meaning of continuity, how close of an approximation is close enough? The key to the answer is that a real number may be approximated as precisely as you wish, as long as you put enough work into it. So for f to be continuous at c, we should be able to demand that $f(x)$ and $f(c)$ be as close together as we like (as long as we still allow for a positive distance between them). But in order to achieve that result, it's fair in turn to demand that x be as close to c as necessary (again as long as we still allow the distance to be positive). The distance between two numbers is given by subtracting and taking the absolute value, so we need to be able to ensure that $|f(x) - f(c)|$ is as small as we want (but positive) by making $|x - c|$ as small as we need (but positive).

The traditional symbols for these small but positive distances are the Greek letters 'ϵ' (lowercase Epsilon) and 'δ' (lowercase Delta). For this reason, this is sometimes called the ϵ-δ (or epsilon-delta) definition; this general method of designing definitions and proving theorems is also called *epsilontics*. So here is the rigorous definition:

> A function f is **continuous** at a real number c if $f(c)$ exists and, for each positive number ϵ (no matter how small), there is some positive number δ (possibly quite small), such that whenever $|x - c| < \delta$ and $f(x)$ exists, $|f(x) - f(c)| < \epsilon$.

This is fairly complicated, but you can view it as a game, involving a function f and a number c such that $f(c)$ exists.

- I challenge you with a positive number ϵ.
- You respond with a positive number δ.
- I reply with a value of x such that $|x - c| < \delta$ and $f(x)$ exists.
- You win if $|f(x) - f(c)| < \epsilon$.

If you can win this game, no matter what choices I make, then f is continuous at c. On the other hand, if I can win no matter what choices you make, then f has a discontinuity at c.

To see how this matters in practice, suppose again that $f(x) = x^2$ for all x and you're told that $x \approx 5$; you want to judge how precisely you know that $x^2 \approx 25$. To be specific, suppose that you want to be guaranteed that x^2 rounds to 25 to at least 2 digits after the decimal point, in other words that $|x^2 - 25| < \frac{1}{2} \times 10^{-2}$. (That is, ϵ is $\frac{1}{2} \times 10^{-2} = 0.005$.) This means that you want x^2 to be between $25 - \frac{1}{2} \times 10^{-2} = 24.995$ and $25 + \frac{1}{2} \times 10^{-2} = 25.005$. Taking square roots (and assuming that x is positive, since it's near 5), this means that x is between $\sqrt{24.995} \approx 4.999\,499\,97$ and $\sqrt{25.0005} \approx 5.000\,499\,98$. To be really sure that this is true, round up the lower number and round down the upper number: x should be between 4.9995 and 5.0004. Notice that $|4.9995 - 5| = 0.0005$, while $|5.0004 - 5| = 0.0004$. So to ensure that x lies in the correct interval, we should insist that $|x - 5| < 0.0004$. (That is, δ is 0.0004; when you have a choice, always use the smaller value of δ to be safe.) So if you can verify that x is at least *that* close to 5, then you can be confident that x^2 is at least as close to 25 as you need. (That f is continuous at 5 means that no matter how precisely you need to know that $x^2 \approx 25$, you'll be able to perform a calculation like this, at least in principle, to find out how precisely you need to require that $x \approx 5$.)

Here are a few more definitions to round out the topic; in all of these definitions, f is a function, c is a number, and S is a set (of numbers).

- f is **left-continuous at** c (or *continuous at c from the left* or *from below*) if the function $(x \mapsto f(x) \text{ for } x \leq c)$ (that is the same as f on the interval $(-\infty, c]$ but undefined on the interval (c, ∞)) is continuous at c.
- f is **right-continuous at** c (or *continuous at c from the right* or *from above*) if the function $(x \mapsto f(x) \text{ for } x \geq c)$ is continuous at c.
- f is **continuous on** S if f is continuous at c whenever $c \in S$ (so in particular, f must be defined on S).
- f is just plain **continuous** if f is continuous on its domain (so continuous at every number where it is defined).

Left and right continuity will not come up much, although sometimes it is useful to know that f is continuous at c if and only if it is both left-continuous and right-continuous there.

However, the other two definitions above will be used often. It will be especially common to say that a function is continuous on an open interval (a, b). This means that we don't care whether it's defined at numbers less than a or greater than b (or even at a or b themselves, since the interval is open) or (even if it is defined there) whether it's continuous there; we only care about what is happening between a and b.

2.3 *Directions*

A **direction** in some variable describes not only whether the variable is increasing or decreasing (that is its literal direction on a number line) but also if there is a limiting value that it approaches but does not reach. The basic directions that we study in this course take the following four forms, where x may be any variable and c may be any constant:

- $x \to \infty$: as x increases without bound (or as x approaches positive infinity);
- $x \to -\infty$: as x decreases without bound (or as x approaches negative infinity);
- $x \to c^-$: as x increases towards c (or as x approaches c from the left, or from below);
- $x \to c^+$: as x decreases towards c (or as x approaches c from the right, or from above).

Any two or more of these directions may be combined, but the only type of combined direction in the textbook is this:

- $x \to c$: as x approaches c (from either direction, or even both at once, jumping back and forth);

which is the combination of $x \to c^-$ and $x \to c^+$. That said, other combinations are also sometimes studied, especially the combination of $x \to \infty$ and $x \to -\infty$, which is written $x \to \pm\infty$: as x approaches positive or negative infinity. (You can also consider fancier directions, for example as x increases without bound *while taking only integer values*, which is relevant to the material in Section 9.1 of the textbook and which I will get to in Chapter 6. For now, however, I'll stick to combinations of the types of directions relevant to Chapter 2.)

It's sometimes convenient to think of ∞ and $-\infty$ as numbers like the real number c, only numbers of an infinite magnitude. Similarly, it's sometimes convenient to think of c^+ and c^- as numbers that are infinitely close to (but distinct from) the real number c. Then the meanings of the directions are as follows:

- $x \to \infty$: what happens when x is positive and infinite?
- $x \to -\infty$: what happens when x is negative and infinite?
- $x \to c^-$: what happens when x is infinitely close to but less than c?
- $x \to c^+$: what happens when x is infinitely close to but greater than c?
- $x \to c$: what happens when x is infinitely close to but distinct from c?
- $x \to \pm\infty$: what happens when the absolute value of x is infinite?

This can be made rigorous, by extending the real number system to the *hyperreal* number system, although I will not talk about how to do that in these notes. But in any case, it can be useful for intuition.

Ultimately, the important thing about a direction is what happens *eventually* as you move in that direction. So for example, as $x \to \infty$, it is eventually true that $x > 0$, that $x > 1$, that $x > 2$, and so on. Besides that ... well, that's it, really. If any statement P is true as $x \to \infty$, then it's true because there is some fixed number M (which you may assume is a whole number, although you don't have to do this) such that P is true whenever $x > M$. For example, $x^2 > 4$ as $x \to \infty$, because $x^2 > 4$ whenever $x > 2$. (It's also true that $x^2 > 4$ whenever $x < -2$, but that's irrelevant to what happens as $x \to \infty$.)

Similarly, P is true (eventually) as $x \to -\infty$ if there is some number M such that P is true whenever $x < -M$. Also, P is true in the combined direction $x \to \pm\infty$ if it is true both as $x \to \infty$ and as $x \to -\infty$, in other words if there is some number M such that P is true whenever $|x| > M$. Next, P is true as $x \to c^+$ if there is some positive number δ (which you may assume is $1/M$ for some natural number M, although you don't have to do this) such that P is true whenever $c < x < c + \delta$; and P is true as $x \to c^-$ if there is some positive number δ such that P is true whenever $c - \delta < x < c$. Finally, P is true as $x \to c$ if it is true both as $x \to c^+$ and as $x \to c^-$, in other words if there is some positive number δ such that P is true whenever $c - \delta < x < c + \delta$ but $x \neq c$ (or equivalently whenever $0 < |x - c| < \delta$).

For example, $x - 2 \neq 0$ as $x \to 2$, precisely because of the $x \neq 2$ bit; the point of $x \to 2$ is that x is *close* to 2 but still *distinct* from 2. You can't say that $x - 2 > 0$ as $x \to 2$, but at least $(x-2)^2 > 0$; also, $x - 2 > 0$ as $x \to 2^+$. This sort of analysis allows you to simplify things as you work in particular directions. (For example, if you're considering $x \to 2^+$ and see $|x - 2|$, then you can simplify that to $x - 2$, since $x - 2 > 0$ as $x \to 2^+$.)

2.4 *Limits*

If D is any direction and u is any variable quantity, then we indicate the value to which u approaches as change occurs in the indicated direction as

$$\lim_{D} u$$

in a displayed equation or as $\lim_D u$ in running text. (The textbook likes to write u as $f(x)$, and this is certainly convenient when it comes to the formal definition, but in practice you'll start with an expression involving the variable x, and it's not necessary to think of this as given by a function.) We will examine the case when u approaches a real value L, as well as the case when u increases without bound or decreases without bound. In the first case, we say that the limit **converges**; in the second case, we say that the limit **diverges** to (positive or negative) infinity. Other types of behaviour are also possible, which are also kinds of divergence, but I won't try to analyse those now.

A limit as $x \to c$ exists (as one of the three kinds of results that we are considering) if and only if the limits as $x \to c^+$ and as $x \to c^-$ are both this same result. So in total, there are fifteen kinds of limits that we will consider, for the five kinds of directions (four basic and one combined) and the three kinds of results:

$$\begin{array}{lll}
\lim_{x\to\infty} u = L; & \lim_{x\to\infty} u = \infty; & \lim_{x\to\infty} u = -\infty; \\
\lim_{x\to-\infty} u = L; & \lim_{x\to-\infty} u = \infty; & \lim_{x\to-\infty} u = -\infty; \\
\lim_{x\to c^-} u = L; & \lim_{x\to c^-} u = \infty; & \lim_{x\to c^-} u = -\infty; \\
\lim_{x\to c^+} u = L; & \lim_{x\to c^+} u = \infty; & \lim_{x\to c^+} u = -\infty; \\
\lim_{x\to c} u = L; & \lim_{x\to c} u = \infty; & \lim_{x\to c} u = -\infty.
\end{array}$$

To see how to read these aloud, I'll consider the last one as an example; this says that the **limit** of u, as x approaches c, is negative infinity.

If you think of ∞ and $-\infty$ as numbers of an infinite magnitude, then the meanings of the results are as follows:

- $\lim_D u = \infty$: u is positive and infinite;
- $\lim_D u = -\infty$: u is negative and infinite;
- $\lim_D u = L$: u is infinitely close to L (which, unlike the direction $u \to L$, includes being equal to L as a special case).

This can be made into a rigorous definition of limits using the hyperreal number system, but I will only use it for intuition.

There are some alternative notations for limits that are worth knowing. First of all, instead of writing $\lim_D u$, you can also write $u|_D$, analogous to evaluation notation (page 6). That is, $u|_{x=c}$ means whatever u *equals* when x *equals* c, while $u|_{x \to c}$ means whatever u *approaches* (or equals) when x *approaches* (but is still distinct from) c.

The point of a continuous function is that these are the same; that is, f is continuous at c if and only if $f(x)|_{x=c}$ and $f(x)|_{x \to c}$ both exist and are equal. Of course, instead of writing $f(x)|_{x=c}$, you could just write $f(c)$; similarly, instead of writing $f(x)|_{x \to c}$, there is yet another notation for this:

$$f(c^{\pm}) = f(x)|_{x \to c} = \lim_{x \to c} f(x).$$

You can read this as 'f of c plus or minus'; the idea behind 'plus or minus' here is the same as in the English phrase 'more or less', meaning 'approximately', because we're looking at values of f *near* c rather than *at* c. Then f is continuous at c if and only if $f(c^{\pm}) = f(c)$ (including the requirement that these both exist).

The analogous notations for the other types of directions are $f(c^-)$, $f(c^+)$, $f(\infty)$, and $f(-\infty)$. Since things like c^+ and ∞ aren't real numbers, there should be no confusion between this function-limit notation and the usual function-evaluation notation $f(c)$. Since none of these alternative notations for limits are in the textbook, I won't use them very much, but they are good to know; they are short and handy, and you may see them elsewhere.

2.5 *Defining limits*

The simplest type of limit to define is $\lim_{x \to c} f(x)$. Note that this just depends on the function f and the real number c, which is especially clear using the notation $f(c^{\pm})$ in the previous paragraph above. If f is continuous at c, then this is supposed to be $f(c)$. But what if f is undefined at c or has a discontinuity there?

Given a real number L, let $f_{c \mapsto L}$ be the piecewise-defined function given by

$$f_{c \mapsto L}(x) = \begin{cases} f(x) & \text{for } x \neq c, \\ L & \text{for } x = c. \end{cases}$$

That is, $f_{c \mapsto L}$ is almost the same function as f, except that $f_{c \mapsto L}(c) = L$, regardless of what $f(c)$ is (or even whether $f(c)$ exists in the first place). Now here is the definition of the limit:

If there is a unique real number L such that $f_{c \mapsto L}$ is continuous at c, then L is $f(c^{\pm})$.

Note that the limit is undefined if either there is no L that makes $f_{c \mapsto L}$ continuous or if there is more than one L that makes it continuous. But that second possibility is very rare; it only happens if f is undefined approaching c, that is if f is not defined anywhere near c (in which case $f_{c \mapsto L}$ is continuous at c no matter what L is, because there is nothing nearby to compare to).

What if the limit is some kind of infinity? We can't talk about $f_{c \mapsto \infty}$, because then $f_{c \mapsto \infty}(c)$ would have to be ∞, which is not a real number. However, if $f(x)$ is increasing without bound, then $1/f(x)$ should be approaching 0. This *almost* allows us to define when the limit is ∞; the only problem is that $1/f(x)$ still approaches 0 even if $f(x)$ *decreases* without bound as well. Still we can say that

$$\lim_{x \to c} f(x) = \pm\infty \text{ if and only if } \lim_{x \to c} \left(\frac{1}{f(x)} \right) = 0.$$

To finish the definitions that we want, we need to specify the sign of $f(x)$ as well:

$$\lim_{x \to c} f(x) = \infty \text{ if and only if } \lim_{x \to c} \left(\frac{1}{f(x)} \right) = 0 \text{ and } f(x) > 0 \text{ as } x \to c;$$

$$\lim_{x \to c} f(x) = -\infty \text{ if and only if } \lim_{x \to c} \left(\frac{1}{f(x)} \right) = 0 \text{ and } f(x) < 0 \text{ as } x \to c.$$

You an also define things like $\lim_{x \to c} u = L^-$ and $\lim_{x \to c} u = L^+$ by similar restrictions, but I won't be doing that.

Finally, for the general defintion of $\lim_D u$, where D is any direction and u is any expression, suppose (like I did back on page 6) that x and u are both functions of some independent variable t, where x is the variable that appears in the direction D. To be precise, suppose that $u = f(t)$ and $x = g(t)$. If the direction D consists of some additional condition on the variable x, then assume that this condition holds for every value of the function g. (So for $x \to c^-$, suppose that $g(t) < c$ always, and for $x \to c^+$, suppose that $g(t) > c$ always; even for $x \to c$, still suppose that $g(t) \neq c$ always.) Then if the limit of $f(t)$ has the same value (a real number L, ∞, or $-\infty$) whenever the limit of $g(t)$ is the value given by the direction D (a real number c, ∞, or $-\infty$), then that value for the limit of $u = f(t)$ is the limit $\lim_D u$.

(This definition covers much more general cases than the textbook's; for example, $\lim_{x \to 0} (\pm x) = 0$, because whenever $f(t) = \pm g(t)$ and $\lim_D g(t) = 0$, then $\lim_D f(t) = 0$. Intuitively, this should be obvious, since $\pm x \approx 0$ whenever $x \approx 0$, no matter whether it's $+x$ or $-x$. But the textbook's definitions can't make sense of this, technically, since $\pm x$ is not a function of x. The formal definition of the Riemann integral is another case where the textbook technically cannot write it down as a limit but I can.)

The textbook defines limits directly using epsilontics (which is very similar to the epsilontic definition of continuity but slightly more complicated), then defines continuity using limits; I have defined continuity using epsilontics and defined limits using continuity. Our definitions come in different orders, but they are equivalent (at least in the cases where the book gives a defintion at all). In any case, the most important method of calculating limits is this:

If f is continuous at c, then $\lim_{x \to c} f(x) = f(c)$.

This fact makes *most* limits trivial to calculate; but it's the exceptions where all of the interesting stuff happens!

For example, let g be this piecewise-defined function:

$$g(x) = \begin{cases} x + 1 & \text{for } x < 2, \\ x + 3 & \text{for } x \geq 2; \end{cases}$$

consider the limits of $g(x)$ in various directions. Since g is continuous everywhere except at 2, it follows that $\lim_{x \to c} g(x)$ is simply $g(c)$ for every real number c other than 2. There are still a few interesting limits of $g(x)$, however: the limits as $x \to 2^+$, as $x \to 2^-$, as $x \to \infty$, and as $x \to -\infty$. The first of these is $g(2) = 5$, basically because $g(x)$ uses the same formula when $x = 2$ as when $x > 2$; formally, it's because $(x \mapsto x + 3 \text{ for } x \geq 2)$ is continuous. (In other words, g is *right-continuous* at 2.) The next one, the limit as $x \to 2^-$, is 3, even though $g(2) \neq 3$ (so g is *not* left-continuous at 2). But the reason for this limit is essentially the same as the reason for the previous limit; it is that $(x \mapsto x + 1 \text{ for } x \leq 2)$ is continuous. Next, the limit as $x \to \infty$ is ∞, because if x is positive as $1/x \to 0$, then $x + 3$ is positive and $1/(x + 3) \to 0$, or going down to an even more basic level, because $1/(1/t + 3)$ simplifies to $t/(1 + 3t)$, which is continuous as a function of t, positive when t is positive, and 0 when t is 0. Finally, the limit as $x \to -\infty$ is $-\infty$, for essentially the same reason, but now using $1/(1/t + 1)$ and looking at negative values. (This time, $1/(1/t + 1)$ can be positive even when t is negative, but not when t is sufficiently close to 0, which is what matters.)

The analysis in the previous paragraph is somewhat ad hoc, showing how you would work directly from the definitions. The next section is about quick methods, but it will still be useful to think about what happens in various directions.

2.6 *Calculation techniques*

Here I discuss the practical aspects of calculating limits.

The first fact to know about calculating limits is that the limit of the variable itself is already given by the direction:

$$\lim_{x \to c^-} x = c,\ \lim_{x \to c^+} x = c,\ \lim_{x \to c} x = c,\ \lim_{x \to \infty} x = \infty,\ \lim_{x \to -\infty} x = -\infty.$$

A similarly important principle is that the limit of a constant, in *any* direction, is that constant:

$$\lim_D K = K \text{ if } K \text{ is constant.}$$

Of course, we rarely bother with limits as simple as these! However, we have the powerful principle that if an expression is built using only the usual operations,* then the limit of the expression may be computed using these operations.

Explicitly, each of these equations is true whenever the right-hand side is defined (so that in particular the left-hand side is automatically also defined), so long as n is constant and $\lim_D w$ is positive:

$$\begin{aligned}
\lim_D (u+v) &= \lim_D u + \lim_D v; & \lim_D (u-v) &= \lim_D u - \lim_D v; \\
\lim_D (uv) &= \lim_D u \cdot \lim_D v; & \lim_D (u/v) &= \frac{\lim_D u}{\lim_D v}; \\
\lim_D (-u) &= -\lim_D u; & \lim_D (|u|) &= \left|\lim_D u\right|; \\
\lim_D (u^n) &= \left(\lim_D u\right)^n; & \lim_D (w^u) &= \left(\lim_D w\right)^{\lim_D u}; \\
\lim_D (\sqrt[n]{u}) &= \sqrt[n]{\lim_D u}; & \lim_D (\log_v u) &= \log_{\lim_D v} \left(\lim_D u\right); \\
\lim_D (\sin u) &= \sin\left(\lim_D u\right); & \lim_D (\cos u) &= \cos\left(\lim_D u\right); \\
\lim_D (\tan u) &= \tan\left(\lim_D u\right); & \lim_D (\cot u) &= \cot\left(\lim_D u\right); \\
\lim_D (\sec u) &= \sec\left(\lim_D u\right); & \lim_D (\csc u) &= \csc\left(\lim_D u\right); \\
\lim_D (\operatorname{asin} u) &= \operatorname{asin}\left(\lim_D u\right); & \lim_D (\operatorname{acos} u) &= \operatorname{acos}\left(\lim_D u\right); \\
\lim_D (\operatorname{atan} u) &= \operatorname{atan}\left(\lim_D u\right); & \lim_D (\operatorname{acot} u) &= \operatorname{acot}\left(\lim_D u\right); \\
\lim_D (\operatorname{asec} u) &= \operatorname{asec}\left(\lim_D u\right); & \lim_D (\operatorname{acsc} u) &= \operatorname{acsc}\left(\lim_D u\right).
\end{aligned}$$

In this way, we can evaluate most limits.

We can do even more limits if we extend arithmetic to the values $\pm\infty$ as follows, where a is (in general) any real number or $\pm\infty$:

$$\begin{aligned}
a + \infty = \infty + a &= \infty \text{ if } a > -\infty; & a - \infty = -\infty + a &= -\infty \text{ if } a < \infty; \\
a \cdot \infty = \infty \cdot a &= \infty \text{ if } a > 0; & a \cdot \infty = \infty \cdot a &= -\infty \text{ if } a < 0; \\
-\infty \cdot a &= -(\infty \cdot a); & a \div \pm\infty &= 0 \text{ if } -\infty < a < \infty; \\
\infty^a &= \infty \text{ if } a > 0; & (\pm\infty)^a &= 0 \text{ if } a < 0; \\
a^\infty &= \infty \text{ if } a > 1; & a^\infty &= 0 \text{ if } -1 < a < 1; \\
a^{-\infty} &= 0 \text{ if } |a| > 1; & a^{-\infty} &= \infty \text{ if } 0 \le a < 1; \\
\sqrt[a]{\infty} &= \infty \text{ if } 0 < a < \infty; & \sqrt[\infty]{a} &= 1 \text{ if } 0 < a < \infty.
\end{aligned}$$

---

* Addition, subtraction, multiplication, division, absolute values, opposites, reciprocals, raising to powers when the exponent is constant or the base is always positive, extracting roots when the index is constant or the base is always positive, logarithms, trigonometric operations, and inverse trigonometric operations, the same as the list of continuous operations spanning pages 9 and 10

Rather than memorizing all of these, it is usually enough to think to yourself what happens if a given number becomes arbitrarily large.

Finally, we can even divide by zero sometimes, *if* we are computing limits!

$$\lim_D (u/v) = \infty \text{ if } \lim_D u > 0,\ \lim_D v = 0,\ \text{and } v > 0 \text{ in the direction } D;$$
$$\lim_D (u/v) = -\infty \text{ if } \lim_D u > 0,\ \lim_D v = 0,\ \text{and } v < 0 \text{ in the direction } D;$$
$$\lim_D (u/v) = -\infty \text{ if } \lim_D u < 0,\ \lim_D v = 0,\ \text{and } v > 0 \text{ in the direction } D;$$
$$\lim_D (u/v) = \infty \text{ if } \lim_D u < 0,\ \lim_D v = 0,\ \text{and } v < 0 \text{ in the direction } D.$$

In other words, if $v \to 0$ with a consistent sign, then the limit of u/v is plus or minus infinity, depending on how the sign of v compares to the sign of u, as long as u approaches something other than 0.

However, this tells us nothing if $u \to 0$ too; in other words, if you work out the limit as far as $0/0$. The same goes for expressions involving infinity such as $\infty - \infty$, $0 \cdot \infty$, $\infty \div \infty$, ∞^0, and 1^∞, none of which is handled by the rules on the previous page. These are all called **indeterminate forms**. Additionally, the rule for $\lim_D (w^u)$ requires that $\lim_D w > 0$; but even if $w > 0$ in the direction D, it's still possible to have $\lim_D w = 0$. In this case, it's best to look at $1/w$ (whose limit is infinite) instead, but the form 0^0 cannot be treated in this way, so this is *also* an indeterminate form.

To handle an indeterminate form, people typically use an advanced technique such as L'Hôpital's Rule (Section 4.5) or expansion into Taylor series (Chapter 8). However, you can often manipulate the expression algebraically to get something that works. This is especially common for $0/0$, which can usually be simplified by factoring. Specifically, try to factor out $x - c$ from both top and bottom if you get $0/0$ while taking a limit $x \to c$ (or $x \to c^+$ or $x \to c^-$). Or if you get $\pm\infty/\infty$ when taking a limit as $x \to \pm\infty$, divide top and bottom by the largest powers of x that appear in either of them.

While I'm at it, here is another rule, called the Chain Rule for limits. I would *like* to say that $\lim_D f(u) = \lim_{v \to L} f(v)$ whenever $L = \lim_D u$, but this is ***not*** always true! However, this *is* true in some special cases: if f is continuous at L, or if $u \neq L$ in the direction D. If either of these conditions holds, then the rule is valid. Much like the Chain Rule for derivatives and differentials in Section 3.4 and 3.6 below, this is not something that you'll use *directly* if you have all of the other rules in this section, but you might need it in a more theoretical situation where you don't know what function f is.

2.7 *Transcendental limits*

When working with complicated expressions, it sometimes helps to use the **Squeeze Theorem** or **Sandwich Theorem**:

> If $u \leq v \leq w$ in the direction D and $\lim_D u$ and $\lim_D w$ both exist (as L, ∞, or $-\infty$) and are equal, then $\lim_D v$ also exists and equals them too.

For example, $-x \leq |x| \leq x$ always, and the limits of $-x$ and x as $x \to 0$ are both zero, so $\lim_{x \to 0} |x| = 0$ too.

That particular example would be easy to establish directly; much less obvious is $\lim_{x \to 0} \left(\frac{\sin x}{x} \right)$. However, you can see that

$$\cos x \leq \frac{\sin x}{x} \leq 1,$$

at least when $-\pi/2 \leq x \leq \pi/2$, using a geometric argument that is essentially given in the textbook on page 82 at the end of Section 2.4. (Actually, this is even true with $<$ in place of $\leq$, although it's not necessary to check that.) Since this inequality is true as $x \to 0$ and the limits of $\cos x$ and 1 as $x \to 0$ are both 1, this means that $\lim_{x \to 0} \left(\frac{\sin x}{x} \right) = 1$ too.

Unfortunately, there is no general method (as far as I know) that will tell you, once you see $\frac{\sin x}{x}$, what particular compound inequality will help with the Squeeze Theorem. However, once you know the general fact of this limit, you can find many other limits involving trigonometric expressions. For example, given $(\cos x - 1)/x$, you can 'sinusize' (instead of 'rationalize') the numerator by multiplying top and bottom by the conjugate $\cos x + 1$ and using $\cos^2 x = 1 - \sin^2 x$, and this is a technique that will work in general. (The result is that

$$\lim_{x \to 0} \left(\frac{\cos x - 1}{x} \right) = \lim_{x \to 0} \left(-\frac{\sin^2 x}{x(\cos x + 1)} \right) = -\lim_{x \to 0} \left(\frac{\sin x}{x} \right) \lim_{x \to 0} \left(\frac{\sin x}{\cos x + 1} \right) = -(1) \frac{\sin(0)}{\cos(0) + 1} = 0,$$

which the textbook calculuates by a less systematic method on page 83 in Example 5.A of Section 2.4.)

2.8 *Uniform continuity*

The point of a continuous function is that if you only know the value of x approximately, then you can find the value of $f(x)$ approximately. When I made this idea precise using epsilontics, I treated each number in the domain separately, or if you prefer, I treated each point on the graph separately. This definition of continuity is sometimes called *pointwise* continuity for emphasis. However, there is another way to define continuity, called *uniform* continuity because the entire domain (or at least an entire set within the domain) is treated at once. (This section is optional, but it can clarify the meaning of continuity, and it would be necessary in a fully rigorous development that proved all of the theorems.)

Look back at the rough definition of continuity on page 9 in Section 2.1. That was for continuity at c, so let me rewrite it to mean continuity everywhere in the domain of f:

> A function f is **pointwise continuous** if, whever $f(c)$ exists and, whenever $x \approx c$ and $f(x)$ exists, $f(x) \approx f(c)$.

Now let me rephrase that slightly:

> A function f is **uniformly continuous** if, whenever $x \approx y$ and $f(x)$ and $f(y)$ exist, $f(x) \approx f(y)$.

All that I did, besides changing the name of c to y, was to move the hypothesis that f is defined there to a later position in the list of hypotheses. That shouldn't make any difference; however, it changes the emphasis, and the definition isn't fully precise yet.

So now look back at the precise definition on page 10 in Section 2.2. Again, I'll rewrite it to mean continuity everywhere in the domain of f:

> A function f is **pointwise continuous** if, whenever $f(c)$ exists, for each positive number ϵ (no matter how small), there is some positive number δ (possibly quite small), such that whenever $|x - c| < \delta$ and $f(x)$ exists, $|f(x) - f(c)| < \epsilon$.

Now here is the version with c renamed to y and the condition that f is defined there moved later:

> A function f is **uniformly continuous** if, for each positive number ϵ (no matter how small), there is some positive number δ (possibly quite small), such that whenever $|x - y| < \delta$ and $f(x)$ and $f(y)$ exist, $|f(x) - f(y)| < \epsilon$.

Now something important has changed: the number y, just like x was, is only introduced *after* the positive distance δ is chosen. So for pointwise continuity, you can use both ϵ and c to help you choose an appropriate δ; but for uniform continuity, you must choose δ using only ϵ to help you. In other words, the same δ must work everywhere, which is why this notion of continuity is called *uniform*.

Here is the description of pointwise continuity as a game, for a given function f:

- I challenge you with a positive number ϵ and a value of c such that $f(c)$ exists.
- You respond with a positive number δ.
- I reply with a value of x such that $|x - c| < \delta$ and $f(x)$ exists.
- You win if $|f(x) - f(c)| < \epsilon$.

And here is the game of uniform continuity:

- I challenge you with a positive number ϵ.
- You respond with a positive number δ.
- I reply with a value of x and a value of y such that $|x - y| < \delta$ and $f(x)$ and $f(y)$ exist.
- You win if $|f(x) - f(y)| < \epsilon$.

This game is harder to win, because you have less information when it is your turn to choose a number.

Thus, every uniformly continuous function is pointwise continuous (that is continuous wherever it is defined), but not every pointwise continuous function is uniformly continuous. Two good examples are $f(x) = x^2$ and $f(x) = 1/x$. In the case of x^2, as $c \to \infty$, you will need to use smaller and smaller values of δ for any given value of ϵ, and no single value of δ will work everywhere. The same is true for $1/x$ but now as $c \to 0^+$. In general, if a continuous function is not uniformly continuous, there will be some direction in which the uniform continuity fails. That is, pointwise continuity fails at specific *points* of discontinuity, while uniform continuity fails at specific *directions* of discontinuity. Or you can think that both of them fail in directions, but with a pointwise discontinuity, the direction is approaching somewhere in the function's domain, while with a uniform discontinuity, the direction may be approaching somewhere outside of the function's domain (possibly even some kind of infinity).

Another way to look at this involves infinitely small and infinitely large numbers. Using these, the rough definitions above *are* rigorous if $\approx$ means infinitely close, but we have to specify whether the numbers involved are standard real numbers or hyperreal numbers that are allowed to be infinite or have infinitesimal variations. If they are all standard real numbers, then $\approx$ means $=$ and the definition becomes trivial, so in both cases, x is allowed to be nonstandard. For pointwise continuity, c must be standard, but for uniform continuity, y does *not* need to be standard. This means that uniform continuity is a stronger requirement, because the condition must hold for nonstandard as well as standard numbers.

Consider again the examples x^2 and $1/x$. First, x^2 fails to be uniformly continuous if x and y are infinite; two infinite numbers can be infinitely close without having squares that are infinitely close. Similarly, $1/x$ fails to be uniformly continuous if x and y are infinitesimal (but nonzero); two nonzero numbers can be infinitely close without having reciprocals that are infinitely close. You can work this out with pure Algebra as long as you keep track of which numbers are infinitesimal and which are not, but I won't write out these examples since you won't have to do any problems like that.

2.9 *Theorems about continuous functions*

There are a few useful theorems about functions that are continuous on a compact interval $[a, b]$. Really, these theorems are about *uniformly* continuous functions (as in the previous section). But the first useful theorem about these is that a continuous function on a compact interval is always uniformly continuous!

Besides this, one big theorem is the **Intermediate Value Theorem**, or IVT. This says that if a function f is continuous on $[a, b]$, then it takes every value between $f(a)$ and $f(b)$. More explicitly,

> Suppose that $a < b$ are real numbers, f is a function defined on $[a, b]$ (at least) and f is (uniformly) continuous on $[a, b]$ (at least). If $f(a) < L < f(b)$ or $f(a) > L > f(b)$, then there is at least one number c with $a < c < b$ such that $L = f(c)$.

To find this number c, one method is the *bisection method*: First see if c might be the average of a and b; caclulate $f\left(\frac{a+b}{2}\right)$. If this is L exactly, great, we found c. If not, then it's either too big or too small, and based on that, the theorem tells us that the real value of c is either between a and this average or between the average and b. So take the average of these two numbers and keep going. (It's not vital to make successive guesses *exactly* halfway between; you can round them off a little bit if that's convenient.) In practice, it's often enough to get $f(c) \approx L$ to some level of precision, so just stop when this is achieved. Otherwise, this process gives a collection of values that are too large and a collection that are too small, and from these you construct a cut (as on page 7) that defines the desired value of c.

This is a process that can be carried out regardless of whether the function f is continuous; the Intermediate Value Theorem is a *theorem* because it guarantees that it will be successful. If you just want $f(c) \approx L$ with a maximum absolute error of ϵ, then the definition of uniform continuity on page 17 directly gives you a value of δ such that, once you're looking at an interval whose width is less than δ, you must be close enough. If you want to get $f(c) = L$ exactly, then you can argue that, for any possible $\epsilon > 0$, you figure out the corresponding value of δ, and once you're looking at an interval with that width or less, you have $|f(c) - L| < \epsilon$ from then on. Since this is true for every positive number ϵ, it follows that $|f(c) - L| = 0$, or $f(c) = L$ as desired. So the bisection method is the technique to find c, and the IVT is the theorem that this technique works, as long as the function is continuous.

A similar theorem is the **Extreme Value Theorem**, or EVT. This says that a continuous function f has both a maximum and minimum value on every compact interval. Explicitly,

> Suppose that $a < b$ are real numbers, f is function defined on $[a, b]$ (at least) and f is (uniformly) continuous on $[a, b]$ (at least). Then there is at least one number c with $a \leq c \leq b$ such that $f(c) \leq f(x)$ whenever $a \leq x \leq b$, and is at least one (different) number c with $a \leq c \leq b$ such that $f(c) \geq f(x)$ whenever $a \leq x \leq b$.

We say that f has a minimum *at* the first value of c and a maximum *at* the second value of c, and that the minimum of f *is* the first value of $f(c)$ while the maximum *is* the second value of $f(c)$. Unlike with the IVT, the proof of this theorem is highly non-constructive, meaning that it gives you no practical method to find (or even approximate) the needed values of c. So as a practical matter, we use ideas from Chapter 4, particularly Section 4.8 on optimization, to find extreme values and where they occur.

Incidentally, if a function is *not* continuous (or if you're looking at a domain that is not compact), then while the maximum and minimum might not exist, it's still possible to talk about something almost as good: the supremum and infimum. If a function f is defined on (at least) a set S, then a number M is the **supremum** of f on S if $f(x) \leq M$ whenever $x \in S$ but $f(x) > y$ for some $x \in S$ whenever $y < M$; similarly, a number m is the **infimum** of f on S if $f(x) \geq m$ whenever $x \in S$ but $f(x) < y$ for some $x \in S$ whenever $y > m$. In other words, the value of the function is never larger than the supremum or smaller than the infimum, and the supremum is the smallest number with its property and the infimum is the largest number with its property. For example, if $f(x) = x^2$ and $S = (2, 3)$ (which is not compact), then the supremum is 9 and the infimum is 4, because although f never takes those values within the set S, it gets arbitrarily close to those values without going beyond them.

A maximum is simply a supremum that is actually a value of the function on the set; similarly, a minimum is an infimum that is actually a value. There's no point in asking about where a supremum or infimum may be unless it is actually a maximum or minimum. That said, you can still talk about the direction in which a continuous function *approaches* its supremum or infimum; using the example from the previous paragraph, $f(x)$ approaches the supremum 9 as $x \to 3^-$, and $f(x)$ approaches the infimum 4 as $x \to 2^+$. The nice thing about the supremum and infimum is that they always exist, at least if you allow infinite values. (If a function takes arbitrarily large values on S, then its supremum there is ∞; if it takes arbitrarily small values, then its infimum is $-\infty$. Also technically, if the set S is the empty set, then the supremum of any function on S is $-\infty$, while the infimum is ∞; this is the only way that the infimum can be larger than the supremum.) The point of the EVT is that the supremum and infimum are actually a maximum and minimum (which includes that they must be finite) when S is a compact interval and f is (uniformly) continuous on S.

3 Differentiation

The single most important topic in Calculus is probably *differentiation*. Whereas limits tell us *where* a quantity is going as it changes, differentiation tells us *how quickly* the quantity is changing. Technically, the question answered by limits does come up more often, but it's also trivial to solve in the vast majority of practical cases (when the variable is given by a continuous function); it may not seem that way while you're doing the problems, but that's just because we're focussing on the exceptions. Differentiation, however, is rarely trivial. That said, it is also rarely difficult; you just need to learn the rules.

A word about notation: As I remarked earlier (at the bottom of page 5), when $y = f(x)$, we can write $\mathrm{d}y/\mathrm{d}x = f'(x)$; both sides of the latter equation are notation for a *derivative*, which is one of the things that differentiation produces. The left-hand side means the derivative of y with respect to x, while f' in the right-hand side is a function which is the derivative of the original function f. (This will be explained in Section 3.2 on page 22.) To say that the derivative of f is f' suggests that the derivative is a basic concept, not a combination of anything more complicated, and that is how the textbook approaches derivatives. But the left-hand side suggests that a derivative is a ratio, the result of dividing $\mathrm{d}y$ by $\mathrm{d}x$, and this is how they were originally used. As for $\mathrm{d}y$ and $\mathrm{d}x$ themselves, they are the *differentials* of y and x; a differential is another thing that differentiation produces. (These will be explained in Section 3.6 on page 26.)

I will start with an intuitive description of differentials, then turn to derivatives for a precise definition, then back to differentials to tie it all together. (Then I'll bring up some applications and the like.)

3.1 *Differences*

I'll introduce differentials by starting with a related concept that can be done with pure Algebra. If a variable quantity x changes from the value a to the value b, then the *difference* between these two values is $\Delta x = b - a$. (The triangle here is an uppercase Greek letter Delta, so Δx is often read 'Delta Ex', but you can also pronounce 'Δ' as 'difference' or 'change in'.) More generally, as x changes from a to b, some other quantity y may change as well, although generally between different values. Whatever the difference in those values is, that is the **difference** in y when x is a and Δx is $b - a$, written $\Delta y|_{\substack{x=a,\\ \Delta x=b-a}}$. Or to put it another way, if x changes from a to $a + h$, then y will change between two values, and the difference between these is $\Delta y|_{\substack{x=a,\\ \Delta x=h}}$.

Here are some examples of what this is saying. Suppose that $y = x^2$, and consider what Δy is in various situations.

- What is $\Delta y|_{\substack{x=1,\\ \Delta x=2}}$, or said with more words, what is Δy when $x = 1$ and $\Delta x = 2$? (Or with even more words, how much does $y = x^2$ change when x starts at 1 and increases by 2?) The answer is $(1+2)^2 - (1)^2 = 9 - 1 = 8$. That is, as x changes from 1 to $1 + 2 = 3$, $y = x^2$ changes from 1 to 9, so the change in y is $9 - 1 = 8$.
- What is Δy when $x = 1$, without specifying a value of Δx? Now $\Delta y|_{x=1} = (1 + \Delta x)^2 - (1)^2 = (1 + 2\,\Delta x + \Delta x^2) - 1 = 2\,\Delta x + \Delta x^2$. (The convention on order of operations is that Δx^2 means $(\Delta x)^2$. Note that Δx does *not* mean Δ times x; that is not meaningful.) You can see that if you now put in 2 for Δx, then you get $\Delta y = 2(2) + (2)^2 = 8$, so this agrees with the previous example.
- Finally, what is Δy in general? Now $\Delta y = (x + \Delta x)^2 - x^2 = (x^2 + 2x\,\Delta x + \Delta x^2) - x^2 = 2x\,\Delta x + \Delta x^2$. Again, you can see that if you put in 1 for x, then you get back the previous example.

We can summarize this result as $\Delta(x^2) = 2x\,\Delta x + \Delta x^2$. If we were doing a lot of work with differences, then this would be a basic rule that you would use repeatedly; as it is, our real work will be with differentials, and the rules for them (on page 27) are (mostly) simpler.

You can also divide differences to get a **difference quotient**. For example, if (as before) $y = x^2$, $x = 1$, and $\Delta x = 2$, then

$$\frac{\Delta y}{\Delta x} = \frac{(1+2)^2 - (1^2)}{2} = \frac{8}{2} = 4.$$

You should recognize this as the calculation of a slope or a rate of change. Indeed, if $y = f(x)$ for a fixed function f, then we can say that the **average rate of change** of f from 1 to $1 + 2 = 3$, or the average rate of change of f on the interval $[1, 3]$, is this number $(\Delta y/\Delta x)|_{\substack{x=1,\\ \Delta x=2}} = 4$ that I just calculated. We can again find a general formula without specific values of x or Δx:

$$\frac{\Delta y}{\Delta x} = \frac{(x + \Delta x)^2 - (x^2)}{\Delta x} = \frac{2x\,\Delta x + \Delta x^2}{\Delta x} = 2x + \Delta x.$$

This may be called the **average rate of change** of y with respect to x, in general. To check, if $x = 1$ and $\Delta x = 2$, then this agrees with the result above: $(\Delta y/\Delta x)|_{\substack{x=1,\\ \Delta x=2}} = 2(1) + (2) = 4$.

The idea behind a *differential* is that is an *infinitesimal* (or infinitely small) difference, but not (generally) no difference at all. (I'll worry later about how to make that precise.) In place of the uppercase Greek letter 'Δ' for a standard (not generally infinitesimal) change, we use the lowercase Latin letter 'd' for an infinitesimal change. A variable x might not have a differential, but if it does, then $\mathrm{d}x$ is that differential. Look at the equation $\Delta y = 2x\,\Delta x + \Delta x^2$ (which you'll remember is true when $y = x^2$) and imagine Δx becoming infinitely small; the term Δx^2 matters less and less compared to the other term. So in the end, we'll have $\mathrm{d}y = 2x\,\mathrm{d}x$ and nothing more. It's easier to make this precise by looking at the difference quotient $\Delta y/\Delta x = 2x + \Delta x$, which will become $\mathrm{d}y/\mathrm{d}x = 2x$, so that's where we'll start. In fact, I'll really start with the function $f(x) = x^2$ and see how a new function $f'(x) = 2x$ may be derived from it. But afterwards, I'll come back to $\mathrm{d}x$ and $\mathrm{d}y$.

3.2 *Derivatives of functions*

Given any function f and a number a in the domain of f, the **difference quotient function** of f at a is a function $\tilde{f}_a$, given by

$$\tilde{f}_a(h) = \frac{f(a + h) - f(a)}{h}.$$

(Note that $\tilde{f}_a$ is *not* standard notation; there are many symbols that people use for this, and our textbook gives it no name at all.) In other words, if $y = f(x)$, then $\tilde{f}_a(h)$ is the difference quotient $(\Delta y/\Delta x)|_{\substack{x=a,\\ \Delta x=h}}$. Notice that $\tilde{f}_a$ is not defined at 0, as this would give $0/0$ as the difference quotient. (In general, it's defined at any value h such that $h \neq 0$ and f is defined at $a + h$.)

However, $0/0$ is an indeterminate form that invites us to take a limit. So, the **derivative** of f at a is the limit of $\tilde{f}_a$ approaching 0:

$$f'(a) = \tilde{f}_a(0^{\pm}) = \lim_{h \to 0} \tilde{f}_a(h) = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h}.$$

When this exists, we say that f is **differentiable** at a. This is the definition in the textbook (see page 117 at the end of Section 3.1), except that the book doesn't bother to give a name to the difference quotient function $\tilde{f}_a$.

Because limits are closely related to continuity, it's possible to give a definition of the derivative based on continuity. First, extend the definition of $\tilde{f}_a$ like this:

$$\tilde{f}_a(h) = \begin{cases} \dfrac{f(a + h) - f(a)}{h} & \text{for } h \neq 0, \\ f'(a) & \text{for } h = 0. \end{cases}$$

If there exists a unique number $f'(a)$ that makes this function continuous at 0, then that number is the derivative of f at a; if there isn't, then this derivative doesn't exist and f is not differentiable at a. As it is, this is just the usual definition stated with different terminology. Now I'll do a little algebra on $\tilde{f}_a$: if $h \neq 0$ and f is defined at $a+h$, then

$$\begin{aligned}
\tilde{f}_a(h) &= \frac{f(a+h)-f(a)}{h}, \\
h\,\tilde{f}_a(h) &= f(a+h)-f(a), \\
h\,\tilde{f}_a(h)+f(a) &= f(a+h), \\
f(a+h) &= f(a)+\tilde{f}_a(h)\,h;
\end{aligned}$$

if $h=0$, then this equation is still true as long as $\tilde{f}_a$ is defined at 0, since then it just says that $f(a) = f(a)$. So another way to define the derivative is to say that f is differentiable at a if there exists a function $\tilde{f}_a$ that is continuous at 0 and satisfies the last equation above (for all h such that f is defined at $a+h$), and then $f'(a) = \tilde{f}_a(0)$. One reason that this is useful is that having the entire function $\tilde{f}_a$ can help with proving theorems about derivatives; see the next section.

3.3 *Theorems about derivatives*

Every operation has a corresponding rule for derivatives. To begin with, recall that if f and g are functions, then $f+g$ is another function, which is defined wherever both f and g are defined, and whose values are given by $(f+g)(x) = f(x)+g(x)$. We similarly have $f-g$, fg and f/g (but the last of these is undefined wherever the value of g is zero, even if f and g are both defined there).

The theorems about their derivatives are as follows:

- The Sum Rule: $(f+g)' = f'+g'$,
- The Difference Rule: $(f-g)' = f'-g'$,
- The Product Rule: $(fg)' = f'g+fg'$,
- The Quotient Rule: $(f/g)' = \dfrac{f'g-fg'}{g^2}$.

These are equations about functions; you can also put an argument into them:

$$\begin{aligned}
(f+g)'(x) &= f'(x)+g'(x), \\
(f-g)'(x) &= f'(x)-g'(x), \\
(fg)'(x) &= f'(x)g(x)+f(x)g'(x); \\
(f/g)'(x) &= \frac{f'(x)g(x)-f(x)g'(x)}{g(x)^2}.
\end{aligned}$$

A general strategy to prove these is to apply the equation for $f(a+h)$ from the previous section. For example, to prove that fg is differentiable wherever f and g are, with $(fg)' = f'g+fg'$, I'll use $\tilde{f}_a$ and $\tilde{g}_a$ along with the limit definition of $(fg)'$:

$$\begin{aligned}
(fg)'(a) &= \lim_{h\to 0}\frac{(fg)(a+h)-(fg)(a)}{h} = \lim_{h\to 0}\frac{f(a+h)\,g(a+h)-f(a)\,g(a)}{h} \\
&= \lim_{h\to 0}\frac{\big(f(a)+\tilde{f}_a(h)\,h\big)\big(g(a)+\tilde{g}_a(h)\,h\big)-f(a)\,g(a)}{h} \\
&= \lim_{h\to 0}\frac{f(a)\,g(a)+f(a)\,\tilde{g}_a(h)\,h+\tilde{f}_a(h)\,h\,g(a)+\tilde{f}_a(h)\,h\,\tilde{g}_a(h)\,h-f(a)\,g(a)}{h} \\
&= \lim_{h\to 0}\frac{\tilde{f}_a(h)\,g(a)\,h+f(a)\,\tilde{g}_a(h)\,h+\tilde{f}_a(h)\,\tilde{g}_a(h)\,h^2}{h} \\
&= \lim_{h\to 0}\big(\tilde{f}_a(h)\,g(a)+f(a)\,\tilde{g}_a(h)+\tilde{f}_a(h)\,\tilde{g}_a(h)\,h\big) = \tilde{f}_a(0)\,g(a)+f(a)\,\tilde{g}_a(0)+\tilde{f}_a(0)\,\tilde{g}_a(0)\,0 \\
&= f'(a)\,g(a)+f(a)\,g'(a)+f'(a)\,g'(a)\,0 = f'(a)\,g(a)+f(a)\,g'(a).
\end{aligned}$$

(To evaluate the limit near the end, I need $\tilde{f}_a$ and $\tilde{g}_a$ to be continuous at 0.) I used smaller steps than the textbook does on page 134 after Example 6 in Section 3.3 there (which is why my proof is longer), and I think that my proof is a little more straightforward, without the part where you add and subtract something without knowing yet why it will help.

The derivative of a constant function is the constant zero function; that is, if $f(x) = K$ for all x, where K is some constant, then

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{K - K}{h} = \lim_{h\to 0} \frac{0}{h} = \lim_{h\to 0} 0 = 0.$$

This fact may be called the Constant Rule. Using this, a special case of the Product Rule is the Multiple Rule:

$$(kf)'(x) = kf'(x)$$

if k is a constant. Another useful rule is the Power Rule: If $f(x) = x^n$ for all x, where n is a constant, then

$$f'(x) = nx^{n-1}.$$

(For integer values of n, this may be proved by repeated application of the Product or Quotient Rule, and there is a more complicated argument that applies to other rational values of n; however, a complete proof is easiest after considering exponents and logarithms.)

Using these rules, you can differentiate any polynomial function, or more generally any rational function. For a polynomial, you differentiate term by term (allowed by the Sum Rule), ignoring any constant terms (by the Constant Rule). For each term, you apply the Multiple Rule (to leave any coefficients alone) and the Power Rule (to bring down the exponent as a coefficient and subtract one from that exponent). For example, if $f(x) = 3x^4 - 5x^2 + 2x - 12$, then $f'(x) = 3(4x^{4-1}) - 5(2x^{2-1}) + 2(1x^{1-1}) + 0 = 12x^3 - 10x + 2$. For rational functions, you must also apply the Quotient Rule. There are examples in Section 3.3 of the textbook.

3.4 *The Chain Rule*

One more rule, very important for theoretical purposes, is the Chain Rule. Using this, I'll be able to justify a new notation for derivatives and an even faster way to calculate them, so in the end you won't need to refer to the Chain Rule explicitly. However, we need it first to ensure that the new technique will work!

Here is the Chain Rule in function notation:

If g is differentiable at a and f is differentiable at $g(a)$, then $f \circ g$ is differentiable at a and

$$(f \circ g)'(a) = f'\big(g(a)\big)\, g'(a).$$

Here, $f \circ g$ is the *composite* of f after g, defined by $(f \circ g)(x) = f\big(g(x)\big)$.

I'll prove this using $\tilde{g}_a$ and $\tilde{f}_{g(a)}$:

$$\begin{aligned}
(f \circ g)'(a) &= \lim_{h\to 0} \frac{(f \circ g)(a+h) - (f \circ g)(a)}{h} = \lim_{h\to 0} \frac{f\big(g(a+h)\big) - f\big(g(a)\big)}{h} \\
&= \lim_{h\to 0} \frac{f\big(g(a) + \tilde{g}_a(h)\,h\big) - f\big(g(a)\big)}{h} = \lim_{h\to 0} \frac{f\big(g(a)\big) + \tilde{f}_{g(a)}\big(\tilde{g}_a(h)\,h\big)\,\tilde{g}_a(h)\,h - f\big(g(a)\big)}{h} \\
&= \lim_{h\to 0} \frac{\tilde{f}_{g(a)}\big(\tilde{g}_a(h)\,h\big)\,\tilde{g}_a(h)\,h}{h} = \lim_{h\to 0} \Big(\tilde{f}_{g(a)}\big(\tilde{g}_a(h)\,h\big)\,\tilde{g}_a(h)\Big) \\
&= \tilde{f}_{g(a)}\big(\tilde{g}_a(0)\,0\big)\,\tilde{g}_a(0) = \tilde{f}_{g(a)}\big(g'(a)\,0\big)\,g'(a) = \tilde{f}_{g(a)}(0)\,g'(a) = f'\big(g(a)\big)\,g'(a).
\end{aligned}$$

This proof is as straightforward as something so abstract can be, and it can be done immediately and rigorously without postponing things as the textbook does. I have the definition of derivative using $\tilde{f}_a$ to thank for this; this definition of derivative will be handy for some other proofs later on, such as for the Mean-Value Theorem.

One immediately useful consequence of the Chain Rule is a generalized form of the Power Rule (what the textbook calls the Power Chain Rule): If g is differentiable at a and n is a constant, then g^n is also differentiable at a (where $(g^n)(x)$ is defined as $g(x)^n$), and $(g^n)'(a) = ng(a)^{n-1}g'(a)$. The reason is that g^n is a composite $f \circ g$ where f is the power function given by $f(x) = x^n$.

3.5 *Inverse functions*

Related to composite functions, we can also consider inverse functions. If f is an invertible function, then the inverse of f is traditionally denoted f^{-1}; I will follow this, but be warned that it conflicts with the notation in the Power Chain Rule in the previous paragraph! An alternative notation for an inverse function is f^*, but this symbol also has other meanings.

The point of an inverse function is that $f(a) = b$ if and only if $f^{-1}(b) = a$; that is, the input of f^{-1} is the output of f (and the other way around). So, if f is differentiable at a, then one might hope that f^{-1} would be differentiable at $f(a)$. It's easy to use the Chain Rule to see that the derivative, if it exists, must be $1/f'(a)$; this is because the composite of f and f^{-1} is the identity function I (the function whose input equals its output, which is the special case of the Power Rule in which $n = 1$) on the domain of f. That is,

$$1 = I'(a) = (f^{-1} \circ f)(a) = (f^{-1})'\big(f(a)\big)f'(a),$$

so

$$(f^{-1})'\big(f(a)\big) = \frac{1}{f'(a)}.$$

However, this argument *assumes* that f^{-1} is differentiable, that is that $(f^{-1})'$ is defined. Unlike the proofs for the Chain Rule, Product Rule, etc, I'm not starting with the definition of the derivative and proving that it must exist while I calculate the formula for it. Certainly there are some situations where it cannot exist; if $f'(a) = 0$, then $(f^{-1})'\big(f(a)\big)$ would have to be $1/0$, which does not exist. But so far, all that we have proved when $f'(a) \neq 0$ is that $(f^{-1})'\big(f(a)\big)$ must be $1/f'(a)$ *if* it exists at all.

Technically, it is *possible* that $f'(a) \neq 0$ while $(f^{-1})'(a)$ doesn't exist (even supposing that f is invertible), although the only examples that I know of are very complicated (requiring a piecewise definition with infinitely many ever smaller pieces). However, the **Inverse Function Theorem** guarantees that the inverse is differentiable, as long as a simple condition is met:

> Let f be a function and J an interval. Suppose that f is differentiable on J, with $f' \neq 0$ on J. Then f is invertible on J, and for each number a in J, the derivative of the inverse at $f(a)$ exists and equals $1/f'(a)$.

People like to use this theorem, because you don't have to know beforehand that the function is invertible; that is part of the conclusion of the theorem.

For an example of how to use this, consider the square function $f(x) = x^2$, whose derivative is $f'(x) = 2x$. Now, $2x \neq 0$ if and only if $x \neq 0$, so the largest intervals that we could use for J are $(-\infty, 0)$ and $(0, \infty)$. Let's use $J = (0, \infty)$; then inverse of f on J is the square-root function $g(y) = \sqrt{y}$. Since $f'(a) = 2a$, we get $g'\big(f(a)\big) = 1/(2a)$. Putting $a = \sqrt{y}$, so that $f(a) = \sqrt{y}^2 = y$, this means that $g'(y) = 1/(2\sqrt{y}) = \sqrt{y}/(2y)$, so we have defined the square-root function and found its derivative. (Since $a > 0$, $y = a^2 > 0$ too, so this does *not* define $\sqrt{0} = 0$. You can add that in as a special case, but then the square-root function is *not* differentiable there; if it were, then the derivative would have to be $1/(2 \cdot 0)$, which is impossible.)

Of course, you already know about the square-root function, and we can find its derivative using the Power Rule. However, in a fully rigorous development of Calculus from scratch, this theorem could be used to *prove* the existence of roots and also prove the case of the Power Rule for rational exponents (whereas the Product and Quotient Rules only prove it for integer exponents). You can still prove these without the Inverse Function Theorem, using the ideas in Section 1.5 that I applied to cube roots and using the definition of a derivative as a limit. But that requires more Algebra to write out all of the details, especially if you want to generalize to arbitrary roots.

3.6 *Differentials*

Many calculations in calculus are easier to do using ***differentials***. Furthermore, differentials and the related ***differential forms*** are often used in applications, especially (but not only) to physics. The official textbook covers differentials (in Section 3.11), but incompletely and only in one minor application. It then uses differentials again later (mostly in material for Calculus 2 and 3), but they are useful much earlier. So I will make heavy use of them.

If x is a variable quantity, then $\mathrm{d}x$ is the **differential** of x. You can think of $\mathrm{d}x$ as indicating an infinitely small (infinitesimal) change in the value of x, or (better) the amount by which x changes when an infinitesimal change is made (an infinitely small change in the value of the independent variable t that everything is a function of, as on page 6). A precise definition is in Section 3.8, but you will *not* be tested directly on that; what you need to know is how to *use* differentials.

Note that $\mathrm{d}x$ is *not* d times x, and $\mathrm{d}x$ is also *not* exactly a function of x. Rather, x (being a *variable* quantity) should itself be a function of some other quantity t, and $\mathrm{d}x$ is also a function of a sort; so d is an *operator*: something that turns one function into another function. However, an expression like $u\,\mathrm{d}x$ does involve multiplication: it is u times the differential of x.

We often divide one differential by another; for example, $\mathrm{d}y/\mathrm{d}x$ is the result of dividing the differential of y by the differential of x. The textbook introduces this notation early to stand for the *derivative* of y with respect to x, and indeed it is that; but what the book doesn't tell you is that $\mathrm{d}y/\mathrm{d}x$ literally is $\mathrm{d}y$ divided by $\mathrm{d}x$. Unfortunately, $\mathrm{d}^2y/\mathrm{d}x^2$, the second derivative of y with respect to x, is *not* literally $\mathrm{d}^2y = \mathrm{d}(\mathrm{d}y)$ divided by $\mathrm{d}x^2 = (\mathrm{d}x)^2$; for this reason, I prefer the notation $(\mathrm{d}/\mathrm{d}x)^2y$, meaning $(\mathrm{d}/\mathrm{d}x)(\mathrm{d}/\mathrm{d}x)y = (\mathrm{d}/\mathrm{d}x)(\mathrm{d}y/\mathrm{d}x) = \mathrm{d}(\mathrm{d}y/\mathrm{d}x)/\mathrm{d}x$, for the second derivative.

The most important fact about differentials is this: If f is a differentiable function, then

$$\mathrm{d}\big(f(u)\big) = f'(u)\,\mathrm{d}u.$$

That is, the differential of $f(u)$ equals $f'(u)$ times the differential of u, where f' is the derivative of the function f. This fact not only shows the relationship between differentials and derivatives, but also (because u could be any quantity) it encapsulates the **Chain Rule** in differential form. The Chain Rule is an important principle in Calculus, which is often difficult to learn how to use (see Section 3.4); but with differentials it is easy.

In particular, if $y = f(x)$, then

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}\big(f(x)\big)}{\mathrm{d}x} = \frac{f'(x)\,\mathrm{d}x}{\mathrm{d}x} = f'(x),$$

so $\mathrm{d}y$ divided by $\mathrm{d}x$ really is the derivative. For a better example, suppose that you have discovered (say from the definition as a limit) that the derivative of $f(x) = \sqrt{x}$ is $f'(x) = \sqrt{x}/(2x)$. Then this fact can be expressed in differential form:

$$\mathrm{d}(\sqrt{x}) = \frac{\sqrt{x}\,\mathrm{d}x}{2x} \tag{*}$$

(because $\mathrm{d}(\sqrt{x}) = \mathrm{d}\big(f(x)\big) = f'(x)\,\mathrm{d}x = \frac{\sqrt{x}}{2x}\,\mathrm{d}x$). Conversely, if (by performing a calculation with differentials) you discover the equation (*) above, then you know the derivative of f as well:

$$f'(x) = \frac{\mathrm{d}\big(f(x)\big)}{\mathrm{d}x} = \frac{\mathrm{d}(\sqrt{x})}{\mathrm{d}x} = \frac{\sqrt{x}\,\mathrm{d}x}{2x\,\mathrm{d}x} = \frac{\sqrt{x}}{2x}.$$

Whichever of these facts you discover first, once you know them, you know something even more general:

$$\mathrm{d}(\sqrt{u}) = \frac{\sqrt{u}\,\mathrm{d}u}{2u}.$$

(The power to derive this from equation (*) is the Chain Rule.) The value of this is that u can be any expression whatsoever; for example, if $u = x^2 + 1$, then

$$\mathrm{d}\Big(\sqrt{x^2+1}\Big) = \frac{\sqrt{x^2+1}\,\mathrm{d}(x^2+1)}{2(x^2+1)} = \frac{\sqrt{x^2+1}(2x\,\mathrm{d}x)}{2(x^2+1)} = \frac{x\sqrt{x^2+1}\,\mathrm{d}x}{x^2+1}.$$

So now you have learnt a new derivative, $\mathrm{d}\Big(\sqrt{x^2+1}\Big)\Big/\mathrm{d}x = x\sqrt{x^2+1}\Big/(x^2+1)$ without having to calculate it from scratch or explicitly use the Chain Rule for functions.

3.7 *Rules of differentiation*

Every theorem about derivatives of functions may also be expressed as a theorem about differentials. Here are the most common rules:

- The Constant Rule: $\mathrm{d}(K) = 0$ if K is constant.
- The Sum Rule: $\mathrm{d}(u + v) = \mathrm{d}u + \mathrm{d}v$.
- The Translate Rule: $\mathrm{d}(u + C) = \mathrm{d}u$ if C is constant.
- The Difference Rule: $\mathrm{d}(u - v) = \mathrm{d}u - \mathrm{d}v$.
- The Product Rule: $\mathrm{d}(uv) = v\,\mathrm{d}u + u\,\mathrm{d}v$.
- The Multiple Rule: $\mathrm{d}(ku) = k\,\mathrm{d}u$ if k is constant.
- The Quotient Rule: $\mathrm{d}\left(\frac{u}{v}\right) = \frac{v\,\mathrm{d}u - u\,\mathrm{d}v}{v^2}$.
- The Power Rule: $\mathrm{d}(u^n) = nu^{n-1}\,\mathrm{d}u$ if n is constant.
- The Root Rule: $\mathrm{d}(\sqrt[m]{u}) = \frac{\sqrt[m]{u}\,\mathrm{d}u}{mu}$ if m is constant.

Of these, only the Constant Rule, the Sum Rule, the Product Rule, and the Power Rule are absolutely necessary, since every other expression built out of the operations in the rules above can be built out of the operations in these four rules. However, it is often handy to use all of these rules; it is up to you how many of these rules to learn. (The Power Rule given here really corresponds to the Power *Chain* Rule in the textbook, because it incorporates the Chain Rule within it. Also, the Root Rule is not in the textbook; it's not necessary, because a root can be algebraically transformed into a power, but the version here automatically rationalizes the denominator, which can be convenient.)

In addition, every time that you learn the derivative of a new function, you learn a new rule for differentials, by applying the Chain Rule to that function. I already showed you an example of this on page 26: applying the Chain Rule to the function $f(x) = \sqrt{x}$ gives the special case of the Root Rule for $n = 2$. Here are a few other functions whose derivatives you will learn, expressed as rules for differentials:

- $\mathrm{d}(\exp u) = \exp u\,\mathrm{d}u$.
- $\mathrm{d}(\ln u) = \frac{\mathrm{d}u}{u}$.
- $\mathrm{d}(\sin u) = \cos u\,\mathrm{d}u$.
- $\mathrm{d}(\cos u) = -\sin u\,\mathrm{d}u$.
- $\mathrm{d}(\operatorname{atan} u) = \frac{\mathrm{d}u}{u^2 + 1}$.

And more! (To be clear, $\exp u$ means e^u, $\ln u = \log_{\mathrm{e}} u$, u is in radians in $\sin u$ and $\cos u$, and $\operatorname{atan} u$ is what is also written $\arctan u$, $\operatorname{Tan}^{-1} u$, or $\tan^{-1} u$.)

Notice that each of these rules (in either list) turns the differential on the left into a sum of terms (possibly only one term, or none in the case of the Constant Rule), and each term is an ordinary expression multiplied by a differential (or something algebraically equivalent to this). An expression like this is called a **differential form** (although there also are more general sorts of differential forms). If, when you are calculating the differential of an expression, your result at any stage is *not* like this, then you have made a mistake!

3.8 *Defining differentials*

To formally define what differentials are and prove their properties, I'll make the same assumption that I made at the beginning of these notes (see page 6): there is an independent variable t that every other variable is a function of. Back then, I said that if $u = f(t)$, then $u|_{t=a} = f(a)$. Now I'll say that, if $u = f(t)$ and the function f is differentiable, then

$$\mathrm{d}u|_{\substack{t=a,\\ \mathrm{d}t=h}} = f'(a)\,h.$$

More generally, if $u = f(t)$ and $v = g(t)$, then

$$(u\,\mathrm{d}v)|_{\substack{t=a,\\ \mathrm{d}t=h}} = f(a)\,g'(a)\,h,$$

and similarly for sums of terms like this (or in principle, even for more general expressions).

Again, this is abstract, but since you don't need to consider the specific functions f and g in the definition, concrete applications are straightforward. For example:

$$\begin{aligned}(2x\,\mathrm{d}x + 3\,\mathrm{d}x)|_{\substack{x=4,\\ \mathrm{d}x=0.05}} &= 2(4)(0.05) + 3(0.05) = 0.55,\\ (2x\,\mathrm{d}x + 3y\,\mathrm{d}y)|_{\substack{x=4,y=5,\\ \mathrm{d}x=0.05,\mathrm{d}y=0.02}} &= 2(4)(0.05) + 3(5)(0.02) = 0.7.\end{aligned}$$

(I've put small numbers in for $\mathrm{d}x$ and $\mathrm{d}y$, because this is most often what comes up in practice, although for theoretical purposes it doesn't matter.) It's now more common to be given only partial information; for example:

$$\begin{aligned}(2x\,\mathrm{d}x + 3\,\mathrm{d}x)|_{x=4} &= 2(4)\,\mathrm{d}x + 3\,\mathrm{d}x = 11\,\mathrm{d}x,\\ (2x\,\mathrm{d}x + 3y\,\mathrm{d}y)|_{x=4,y=5} &= 2(4)\,\mathrm{d}x + 3(5)\,\mathrm{d}y = 8\,\mathrm{d}x + 15\,\mathrm{d}y.\end{aligned}$$

Notice that you *don't* plug in the values of x and y inside the differential operator d; if you're not given values of $\mathrm{d}x$ and $\mathrm{d}y$, then those differentials must remain in the answer.

While expressions like the above come up occasionally (such as the discussion of linear approximation in Section 4.2), the main purpose of a precise definition is to prove theorems. (That's how we can be sure that the rules of Calculus will always work, at least when the definitions that prove them can be made to apply.) On the previous page, I gave a list of rules for differentials; you can prove these using the precise definition of differential and the known rules for derivatives of functions. For example, if $u = f(t)$ and $v = g(t)$, then $uv = f(t)\,g(t) = (fg)(t)$. Therefore,

$$\mathrm{d}(uv)|_{\substack{t=a,\\ \mathrm{d}t=h}} = (fg)'(a)\,h = \big(f'(a)\,g(a) + f(a)\,g'(a)\big)\,h = g(a)\,f'(a)\,h + f(a)\,g'(a)\,h = (v\,\mathrm{d}u + u\,\mathrm{d}v)|_{\substack{t=a,\\ \mathrm{d}t=h}}.$$

Here, I've used the formal definition of differential along with the Product Rule for derivatives of functions. The conclusion is that $\mathrm{d}(uv)$ and $v\,\mathrm{d}u + u\,\mathrm{d}v$ always evaluate to the same result, so

$$\mathrm{d}(uv) = v\,\mathrm{d}u + u\,\mathrm{d}v,$$

which is the Product Rule for differentials. In the same way, all of the rules for differentials follow from rules for derivatives of functions.

The Chain Rule is an important special case, so I'll prove it too. If $u = g(t)$ and f is any function, then $f(u) = f(g(t)) = (f \circ g)(t)$, so if f is differentiable, then

$$\mathrm{d}\big(f(u)\big)\big|_{\substack{t=a,\\ \mathrm{d}t=h}} = \mathrm{d}\big((f\circ g)(t)\big)\big|_{\substack{t=a,\\ \mathrm{d}t=h}} = (f\circ g)'(a)\,h = f'\big(g(a)\big)\,g'(a)\,h = \big(f'(u)\,\mathrm{d}u\big)\big|_{\substack{t=a,\\ \mathrm{d}t=h}}.$$

Again, I used the definition of differential and the Chain Rule for functions, and my conclusion is the Chain Rule for differentials:

$$\mathrm{d}\big(f(u)\big) = f'(u)\,\mathrm{d}u$$

whenever f is a differentiable function.

It's not really essential to assume that there exists a *single* independent variable t that every other variable is a function of, and I'll stop making that assumption in Calculus 3 (if you stick around that long). Then the formal definition will become a little trickier, but all of the rules for differentials will continue to apply exactly as I stated them on the previous page.

3.9 *Using differentials*

The main technique for using differentials is simply to take the differential of both sides of an equation. However, you may only do this to an equation that holds *generally*, but *not* to an equation that holds only for *particular* values of the variables. (Ultimately, this is because d is an operator, not a function, so it must be applied to entire functions, not only to particular values of those functions.)

The simplest case is an equation such as $y = \exp(x^2)$, if you want the derivative of y with respect to x. So:

$$\begin{aligned} y &= \exp(x^2); \\ \mathrm{d}y &= \mathrm{d}\big(\exp(x^2)\big) = \exp(x^2)\,\mathrm{d}(x^2) = \exp(x^2)\cdot 2x\,\mathrm{d}x = 2x\exp(x^2)\,\mathrm{d}x; \\ \frac{\mathrm{d}y}{\mathrm{d}x} &= 2x\exp(x^2). \end{aligned}$$

Now you have the derivative. If you want the second derivative, then do this process again:

$$\begin{aligned} \mathrm{d}y/\mathrm{d}x &= 2x\exp(x^2); \\ \mathrm{d}(\mathrm{d}y/\mathrm{d}x) &= \mathrm{d}\big(2x\exp(x^2)\big) = \exp(x^2)\,\mathrm{d}(2x) + 2x\,\mathrm{d}\big(\exp(x^2)\big) \\ &= \exp(x^2)\cdot 2\,\mathrm{d}x + 2x\cdot 2x\exp(x^2)\,\mathrm{d}x = \big(2\exp(x^2) + 4x^2\exp(x^2)\big)\,\mathrm{d}x; \\ (\mathrm{d}/\mathrm{d}x)^2 y = \frac{\mathrm{d}(\mathrm{d}y/\mathrm{d}x)}{\mathrm{d}x} &= 2\exp(x^2) + 4x^2\exp(x^2). \end{aligned}$$

Now we have the second derivative (also written $\mathrm{d}^2y/\mathrm{d}x^2$).

The previous example began with an equation solved for y. But we don't need this; suppose instead that we have $y^5 + x^2 = x^5 + y$ (which *cannot* be solved for either variable using the usual algebraic operations of addition, subtraction, multiplication, division, powers, and roots). Undaunted, we forge ahead anyway:

$$\begin{aligned} y^5 + x^2 &= x^5 + y; \\ \mathrm{d}(y^5 + x^2) &= \mathrm{d}(x^5 + y); \\ \mathrm{d}(y^5) + \mathrm{d}(x^2) &= \mathrm{d}(x^5) + \mathrm{d}y; \\ 5y^{5-1}\,\mathrm{d}y + 2x^{2-1}\,\mathrm{d}x &= 5x^{5-1}\,\mathrm{d}x + \mathrm{d}y; \\ 5y^4\,\mathrm{d}y - \mathrm{d}y &= 5x^4\,\mathrm{d}x - 2x\,\mathrm{d}x; \\ (5y^4 - 1)\,\mathrm{d}y &= (5x^4 - 2x)\,\mathrm{d}x; \\ \frac{\mathrm{d}y}{\mathrm{d}x} &= \frac{5x^4 - 2x}{5y^4 - 1}. \end{aligned}$$

This process is called **implicit differentiation**.

The second derivative is a little more straightforward at first (or it would be if we didn't have to use the Quotient Rule), but there is a twist at the end:

$$\begin{aligned} \mathrm{d}y/\mathrm{d}x &= \frac{5x^4 - 2x}{5y^4 - 1}; \\ \mathrm{d}(\mathrm{d}y/\mathrm{d}x) &= \mathrm{d}\left(\frac{5x^4 - 2x}{5y^4 - 1}\right) = \frac{(5y^4 - 1)\,\mathrm{d}(5x^4 - 2x) - (5x^4 - 2x)\,\mathrm{d}(5y^4 - 1)}{(5y^4 - 1)^2} \\ &= \frac{(5y^4 - 1)(20x^3 - 2)\,\mathrm{d}x - (5x^4 - 2x)(20y^3)\,\mathrm{d}y}{(5y^4 - 1)^2} \\ &= \frac{20x^3 - 2}{5y^4 - 1}\,\mathrm{d}x - \frac{20y^3(5x^4 - 2x)}{(5y^4 - 1)^2}\,\mathrm{d}y; \\ (\mathrm{d}/\mathrm{d}x)^2 y = \frac{\mathrm{d}(\mathrm{d}y/\mathrm{d}x)}{\mathrm{d}x} &= \frac{20x^3 - 2}{5y^4 - 1} - \frac{20y^3(5x^4 - 2x)}{(5y^4 - 1)^2}\,\frac{\mathrm{d}y}{\mathrm{d}x} \\ &= \frac{20x^3 - 2}{5y^4 - 1} - \frac{20y^3(5x^4 - 2x)}{(5y^4 - 1)^2}\,\frac{5x^4 - 2x}{5y^4 - 1} \end{aligned}$$

(which could be simplified further). Notice that I substitute the known expression for $\mathrm{d}y/\mathrm{d}x$ in the last step.

Another handy application of differentials is the case where both quantities x and y may be expressed as functions of some other quantity t. (For the purposes of formal definitions, we always assume that this is possible, but now we're really going to use it.) If we start with the same equation as above, then this will give us an equation relating the derivatives with respect to t:

$$\begin{aligned} y^5 + x^2 &= x^5 + y; \\ \mathrm{d}(y^5 + x^2) &= \mathrm{d}(x^5 + y); \\ \mathrm{d}(y^5) + \mathrm{d}(x^2) &= \mathrm{d}(x^5) + \mathrm{d}y; \\ 5y^{5-1}\,\mathrm{d}y + 2x^{2-1}\,\mathrm{d}x &= 5x^{5-1}\,\mathrm{d}x + \mathrm{d}y; \\ 5y^4\frac{\mathrm{d}y}{\mathrm{d}t} + 2x\frac{\mathrm{d}x}{\mathrm{d}t} &= 5x^4\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\mathrm{d}y}{\mathrm{d}t}. \end{aligned}$$

If we have information about one or both of these derivatives, then this equation will often give us useful information to solve a problem. This situation is called **related rates**, since derivatives can be viewed as rates of change (especially derivatives with respect to time t, although the t in the equation above doesn't have to stand for time).

When we get to integrals, differentials become so useful that even the textbook starts using them, but I'll save that for later.

3.10 *The Implicit Function Theorem*

Suppose you have an equation in two variables, say x and y. For simplicity, let's assume that the right-hand side is a constant (in case it's not, you can always subtract the right-hand side from both sides of the original equation to get an equaivalent equation in which the right-hand side is the constant zero). Then if you differentiate both sides of the equation, and assuming that the operations in it are all differentiable, you'll get something of the form

$$u\,\mathrm{d}x + v\,\mathrm{d}y = 0,$$

where u and v are expressions in the same variables x and y. If you want to find $\mathrm{d}y/\mathrm{d}x$, then you can solve this equation to get

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{u}{v}.$$

Of course, this is undefined if $v = 0$, but hopefully that won't happen, at least not for most values of x and y. For example, if $x^2 + y^2 = 25$, then differentiating this yields $2x\,\mathrm{d}x + 2y\,\mathrm{d}y = 0$, so here $u = 2x$ and $v = 2y$. Then solving for $\mathrm{d}y$ and dividing by $\mathrm{d}x$ gives $\mathrm{d}y/\mathrm{d}x = -u/v = -x/y$.

In such a situation, we'd like the original equation to implicitly define y as a function of x, if not totally, then at least partially, that is for values of x near a given value of x and y. For example, if we start with $x^2 + y^2 = 25$, then solving for y gives $y = \pm\sqrt{25 - x^2}$. This is not a function yet, but if we look for a solution near $(x, y) = (3, 4)$, then we can use $y = \sqrt{25 - x^2}$. (And if we look for a solution near $(x, y) = (3, -4)$, then we can use $y = -\sqrt{25 - x^2}$.) However, if we look for a solution near $(x, y) = (5, 0)$, then there is still no function (at least not a continuous one), because there are both positive and negative values of y at points on the graph of the equation arbitrarily close to $(5, 0)$. And this corresponds to the fact that $\mathrm{d}y/\mathrm{d}x = -x/y$ is undefined when $y = 0$.

The **Implicit-Function Theorem** gives a relationship between the ability to solve for $\mathrm{d}y/\mathrm{d}x$ and the ability to express y as a function of x. It says that if you have an equation in x and y with a constant right-hand side and a particular pair of values $(x, y) = (a, b)$ satisfying the equation, and if you differentiate the equation and get $u\,\mathrm{d}x + v\,\mathrm{d}y = 0$, then as long as u and v are made of the usual operations*, both u and v are defined when $(x, y) = (a, b)$, and v is also non-zero when $(x, y) = (a, b)$, then there is

* or more broadly whenever they're continuous as functions of the two variables x and y, but that's subtler than you might expect and so will be deferred to Calculus 3

an interval I around a and a differentiable function f defined on I such that the original equation holds when $y = f(x)$. And furthermore, in that case $f'(x) = -u/v$ (which is just what you'd expect from solving for $\mathrm{d}y/\mathrm{d}x$).

The proof of this theorem is rather advanced, and I won't try to prove it here. (However, it follows from the theorem for existence of solutions to differential equations in Chapter 6.) But it means that you don't have to *assume* that y is a function of x (or even that x and y are both functions of some other independent variable that might not be the same as x) when you do implicit differentiation and calculate $\mathrm{d}y/\mathrm{d}x$. As long as your calculation yields a result, without dividing by zero or any other funny business, then y *must* be a function of x.

An important special case of this is when we already have one function f and we're looking at the equation $x = f(y)$ (notice the unusual order of the variables). Solving this for y would produce $y = f^{-1}(x)$, where f^{-1} is the *inverse* of the function f (which not all functions have). If you rewrite $x = f(y)$ as

$$x - f(y) = 0,$$

then the Implicit-Function Theorem could apply. Differentiating this gives

$$\mathrm{d}x - f'(y)\,\mathrm{d}y = 0,$$

as long as f is differentiable; solving for $\mathrm{d}y$ and dividing by $\mathrm{d}x$ gives

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{f'(y)}.$$

So the Implicit-Function Theorem does apply as long as f is differentiable and $f'(y) \neq 0$. Specifically, if $f(b) = a$, f' is continuous at b, and $f'(b) \neq 0$, then there exists at least a partial inverse function f^{-1} (that is an inverse of a restriction of f to some smaller domain), defined on some interval around a, and furthermore

$$(f^{-1})'(x) = \frac{1}{f'(y)}.$$

Or since $y = f^{-1}(x)$ and $a = f(b)$, this can be expressed without reference to the original equation: If a function f is continuously differentiable at a number b and its derivative is nonzero at b, then there is a partial inverse function f^{-1} defined near $f(b)$, and f^{-1} is differentiable with

$$(f^{-1})'(x) = \frac{1}{f'\big(f^{-1}(x)\big)}.$$

Or if the domain of f is itself an interval and f is differentiable with a nonzero derivative on the *entire* interval, then f has a *total* inverse function f^{-1}, which is also differentiable with the derivative given by the same formula as above. This is the **Inverse-Function Theorem**.

The Inverse-Function Theorem is used to prove the existence and differentiability of lots of functions. Even the humble square-root function can be defined using it: if $f(x) = x^2$, then $f'(x) = 2x$, which is nonzero on the interval $(0, \infty)$. Therefore there exists an inverse function f^{-1} for the restriction of f to that interval, where we may *define* $\sqrt{x}$ to be $f^{-1}(x)$, and the derivative is automatically

$$(f^{-1})'(x) = \frac{1}{f'\big(f^{-1}(x)\big)} = \frac{1}{f'(\sqrt{x})} = \frac{1}{2\sqrt{x}} = \frac{\sqrt{2}}{2x}.$$

(Having established this, we can extend the square-root function by continuity to 0, so that $\sqrt{0} = 0$ is also defined, but notice that this function is *not* differentiable there.) In this way, all of the root functions can be defined and differentiated, which also proves the Root Rule and the Power Rule for all rational exponents. (The irrational exponenents can be handled by approximating them with rational ones, or more slickly by working with exponential functions as in the textbook.) Even more dramatically, we get the inverse of $f(x) = x^5 + x$. The equation for the inverse function cannot be solved with the usual operations, but since the derivative $f'(x) = 5x^4 + 1 > 0$, this inverse function must exist (and be differentiable to boot).

4 Applications of differentiation

There are many applications of derivatives and differentials, some of which I explore here.

4.1 *Derivatives with respect to time*

Derivatives with respect to time are a major application of Calculus. Here are some examples:

Quantity:	Derivative (with respect to time):	Second derivative:	Third derivative:
Position	Velocity	Acceleration	Jerk
Velocity	Acceleration	Jerk	
Speed	Colloquial acceleration		
Acceleration	Jerk		
Net wealth	Net income		
National debt	National deficit		

Position tells you where something is, while **velocity** tells you how it is moving, that is how its position is changing with time. Velocity is not quite the same thing as **speed**, since velocity keeps track of direction as well. (In this class, most problems involving motion will take place in only one dimension, so there are two directions, represented by positive and negative velocity, while speed is the absolute value of velocity.)

The derivative of velocity with respect to time, in other words the second derivative of position with respect to time, is **acceleration** in the technical sense of this term. On the other hand, the derivative of speed is **colloquial acceleration**, which reflects how the term is used in everyday life. Colloquially, we say that an object is accelerating if its speed increases with time (in other words if it is speeding up) and decelerating if its speed decreases (in other words if it is slowing down). But in the technical sense of the term, if an object is moving in the negative direction and slows down, then its velocity is becoming less negative and more positive, and so its acceleration is positive, even though its colloquial acceleration is negative. (For motion in more than one dimension, it's even possible for the colloquial acceleration to be zero even while the technical acceleration is far from zero; this happens when changing direction while travelling at a constant speed.) Colloquial acceleration is also called *scalar acceleration* or (especially in technical contexts) *tangential component of acceleration*; these terms come from the multidimensional case (and you will understand them if you take Calculus 3).

The time derivative of acceleration (in the technical sense) is **jerk**; that makes jerk the second derivative of velocity and the third derivative of position. Whereas position and velocity can't be directly felt, you feel acceleration as a pressure or absence thereof (a sense of falling or being held or pushed), and a sudden change in that acceleration is a jerk or yank. In engineering, acceleration must be controlled because it can destroy objects by crushing; jerk must be controlled because it can destroy objects by shaking them apart. Even higher derivatives of position are sometimes also studied, although the terminology varies.

Turning to finances, your **net wealth** is the total value of all assets that you own minus the value of all of your debts. (If you owe more than you own, then your net wealth is negative.) This is measured in units of money, such as dollars. Your **net income**, on the other hand, is the total value of everything that you receive (as wages, gifts, and so forth) in a period of time minus the value of your expenses. This is measured in units of money per unit of time, such as dollars per year. In finance, the default unit of time is a year, so you'll often say that someone's income is so many dollars, but this really means so many dollars *per year*. Unlike physical motion, money goes in and comes out in discrete chunks, so the continuous ideas of Calculus are only an approximation, but they can be a good approximation for some purposes.

Turning from personal finances to national, a country's government usually has some debt, called the country's **national debt**, and if the government spends more than it receives from taxes and other revenue, then the difference is the **national deficit**. The debt is the total amount of money owed by the government, while the deficit is the additional amount that has to be borrowed in a given period of time. Again, deficit should really be measured in units of money per unit of time; so if someone says that the the U.S. national deficit is nearly 3 trillion dollars, this really means 3 trillion dollars *per year*. This is the same as 30 trillion dollars per decade (since a decade is 10 years), or 250 billion dollars per month. On the other hand, when they say that the U.S. national debt is almost 30 trillion dollars, then they are saying exactly what they mean; this is the net result of all of the deficits (and occasional surpluses, which are negative deficits) in the past.

4.2 *Linear approximation*

Recall from the top of page 23 that if f is differentiable at a, then

$$f(a+h) = f(a) + \tilde{f}_a(h)\, h$$

for some function $\tilde{f}_a$ that's continuous at 0 (and then $\tilde{f}_a(0)$ is $f'(a)$). Since $\tilde{f}_a$ is continuous at 0, we can say that $\tilde{f}_a(h) \approx \tilde{f}_a(0)$ when $h \approx 0$, or in other words, $\tilde{f}_a(h) \approx f'(a)$ when $h \approx 0$. Putting this approximation in the equation above, we get

$$f(a+h) \approx f(a) + f'(a)\, h$$

when $h \approx 0$. Writing x for $a+h$ (so that $h = x - a$), you can also put this as

$$f(x) \approx f(a) + f'(a)\,(x-a)$$

when $x \approx a$. While the left-hand side could be any differentiable function, the right-hand side is a linear function of x; this function is the **linear approximation** to f near a, or the **linearization** of f near a.

The textbook likes to name this function L; so $f(x) \approx L(x) = f(a) + f'(a)\,(x-a)$. I don't like that name, because which function you get as the linear approximation depends on which function you start with as well as on which number a you look at. So I like to write $L_{f,a}$ for the linearization of f near a:

$$f(x) \approx L_{f,a}(x) = f(a) + f'(a)\,(x-a).$$

This is actually only the beginning of a whole sequence of approximations, each (typically) better than the one before it:

$f(x) \approx f(a)$, a constant, if f is continuous at a;
$f(x) \approx f(a) + f'(a)\,(x-a)$, a linear function of x, if f is differentiable at a;
$f(x) \approx f(a) + f'(a)\,(x-a) + \dfrac{1}{2} f''(a)\,(x-a)^2$, a quadratic function of x, if f is twice differentiable at a;
$f(x) \approx f(a) + f'(a)\,(x-a) + \dfrac{1}{2} f''(a)\,(x-a)^2 + \dfrac{1}{6} f'''(a)\,(x-a)^3$, a cubic function of x,
if f is 3-times differentiable at a;
$\vdots$

(This sequence of approximations is covered in Calculus 2; see Section 9.8 of the textbook or page 59 of these notes.)

It's handy to describe linear approximation in terms of differentials and differences. Recall from Section 3.1 that while the differential $\mathrm{d}x$ represents an infinitesimal (infinitely small) change, the difference Δx represents a standard change that is generally *not* infinitely small. If $y = f(x)$, then the linear approximation of f says that

$$\Delta y|_{\substack{x=a,\\ \Delta x=h}} = f(a+h) - f(a) \approx f(a) + f'(a)\, h - f(a) = f'(a)\, h = \mathrm{d}y|_{\substack{x=a,\\ \mathrm{d}x=h}}.$$

So in the end, linear approximation replaces differences with differentials. Although

$$\Delta y|_{\substack{x=a,\\ \Delta x=h}} \approx \mathrm{d}y|_{\substack{x=a,\\ \mathrm{d}x=h}}$$

is the proper way to put it, often one abbreviates this as

$$\Delta y \approx \mathrm{d}y.$$

(But really this only correct if we also have $\Delta x = \mathrm{d}x$, or at least $\Delta x \approx \mathrm{d}x$, because that difference is also replaced by a differential in the approximation.)

More generally, you can say that an equation involving differentials can be replaced by an approximate equation involving differences. For example, if $x^5 + 2x = y^5 + y$, then $5x^4\,\mathrm{d}x + 2\,\mathrm{d}x = 5y^4\,\mathrm{d}y + \mathrm{d}y$ (by differentiating both sides), so $5x^4\,\Delta x + 2\,\Delta x \approx 5y^4\,\Delta y + \Delta y$. Then if you are looking near the only obvious solution, $(x, y) = (0, 0)$, and you want to know the value of y when $x = 0.3$ (so $\Delta x = 0.3 - 0 = 0.3$), you find $5(0)^4(0.3) + 2(0.3) \approx 5(0)^4\,\Delta y + \Delta y$, so $\Delta y \approx 0.6$; in other words, the new y-value is approximately $0 + 0.6 = 0.6$. (The actual solution to $(0.3)^5 + 2(0.3) = y^5 + y$ is $y|_{x=0.3} \approx 0.55$ to 2 decimal places, but I couldn't do that by hand!)

It can be important to know how far off an approximation might be, and this is basically given by the next term in the sequence of approximations on the previous page. To be specific, the Mean-Value Theorem (from Section 4.4) says that $f(x) - f(a)$ (which is the error in the constant approximation $f(x) \approx f(a)$) cannot be any larger in absolute value than $|x - a|$ times the maximum value that f' takes between x and a; similarly, $f(x) - L_{f,a}(x)$ (which is the error in the linear approximation near a) cannot be any larger in absolute value than $|x - a|^2$ times half the maximum value that f'' takes between x and a. However, the details of why this is so are best saved for the full treatment of the entire sequence of approximations that begins on page 59 of these notes.

4.3 *Newton's Method*

If you want to solve an equation $f(x) = 0$, then the Intermediate Value Theorem (page 18) may give you a way to approximate the solution, but it is usually very inefficient. The Newton–Raphson Method (or simply Newton's Method) is usually much faster, although it doesn't always work. Here, you start with a guess x_0, then replace it with a (hopefully) better guess x_1, and so on. These guesses are computed in turn as follows:

$$\begin{aligned} x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)}, \\ x_2 &= x_1 - \frac{f(x_1)}{f'(x_1)}, \\ x_3 &= x_2 - \frac{f(x_2)}{f'(x_2)}, \\ &\vdots \end{aligned}$$

With any luck, none of these guesses will give $f'(x) = 0$ (which makes the next guess undefined) but eventually one will give $f(x) \approx 0$ to whatever precision you want.

The Newton–Raphson Method is guaranteed to work under certain conditions given by the **Newton–Kantorovich Theorem**: If f is differentiable at a, $f(a)$ and $f'(a)$ are nonzero, f is twice differentiable at least between a and $a - 2f(a)/f'(a)$, and

$$|f''(x)| \leq \frac{|f'(a)|^2}{2\,|f(a)|}$$

whenever x is between a and $a - 2f(a)/f'(a)$, then Newton's Method will give a sequence of values between a and $a - 2f(a)/f'(a)$, and the sequence will converge to a solution of $f(x) = 0$ in the sense that the limit $\lim_{n\to\infty} x_n$ exists and $f\left(\lim_{n\to\infty} x_n\right) = 0$. (See Section 6.1 for details of what sequences are and how they converge.) You will not be tested on this theorem in this course (so you can skip these last two paragraphs if you want), but if you ever need to use Newton's Method in real life, then this can be useful to know.

If you want to approximate the solution to within a maximum error of ϵ, then you usually don't need very many steps, although this depends on how small ϵ is and how closely the conditions of the theorem are met. The inequality above says that there is a number $k \leq 1/2$ such that

$$|f(a)|\,|f''(x)| \Big/ |f'(a)|^2 \leq k$$

whenever x is between a and $a - 2f(a)/f'(a)$. Even if $k = 1/2$, the solution can be found with a maximum error of ϵ after at most $1 + \log_2 |f(a)| - \log_2 |f'(a)| - \log_2 \epsilon$ steps (rounded up). However, if $k < 1/2$, then let δ be $\dfrac{2(1-2k)\,|f(a)|}{k(1-k)\,|f'(a)|}$. After at most $1 + \log_2 |f(a)| - \log_2 |f'(a)| - \log_2 \delta = \log_2 k + \log_2 (1-k) - \log_2 (1-2k)$ steps, you'll have the value with a maximum error of δ; but after that, you'll have it with a maximum error of ϵ after at most $1 + \log_2 (\log_2 \delta - \log_2 \epsilon)$ additional steps (which is much fewer than the $\log_2 \delta - \log_2 \epsilon$ additional steps that you would expect from the previous formula).

4.4 *Advanced theorems*

There are various theorems about derivatives and differentials that should seem obvious if you understand the basic idea, but mathematicians have still proved them just to be safe.

For example, the derivative of a function is supposed to tell us how much the output is changing relative to the input. In particular, if the derivative is positive, then the output should increase when the input increases and decrease when the input decreases; conversely, if the derivative is negative, then the output should decrease when the input increases and decrease when the input increases. The first kind of function is called *increasing* and the other is *decreasing*; there are precise theorems that a function whose derivative somewhere is positive or negative must be increasing or decreasing (repsectively) near there. Conversely, if a function has a local extremum, then the derivative must be either zero or undefined there. This fact is key to optimization (see Section 4.8).

Another group of theorems are the mean-value theorems. The point of a derivative is that it can be approximated by a difference quotient; the mean-value theorems reverse this, and show how a difference quotient must (under some conditions) be equal to a derivative somewhere nearby. All of these theorems consider a function f defined on at least an interval $[a,b]$ (with $a < b$) such that f is continuous on all of $[a,b]$ and differentiable at least between a and b (but possibly not at exactly a or b). Specifically, Rolle's mean-value theorem (usually just called **Rolle's Theorem**) says

> If $f(b) - f(a) = 0$, then $f'(c) = 0$ for some c between a and b.

Then Lagrange's mean-value theorem (the default **Mean-Value Theorem**) says

> In any case, $f'(c) = \dfrac{f(b) - f(a)}{b - a}$ for some c between a and b.

Finally, Cauchy's mean-value theorem says

> If g is another function satisfying the same conditions as f and if furthermore g' is never zero between a and b (although it could be zero or undefined at exactly a or b), then $\dfrac{f'(c)}{g'(c)} = \dfrac{f(b) - f(a)}{g(b) - g(a)}$ for some c between a and b.

In Cauchy's mean-value theorem, I like to think of $f(x)$ as u and $g(x)$ as v, so that the left-hand side is $\mathrm{d}u/\mathrm{d}v$ (evaluated at $x = c$) while the right-hand side is $\Delta u/\Delta v$ (evaluated at $x = a$ and $\Delta x = b - a$). That is, if f and g meet the requirements of the theorem, then there is some number c between a and b such that

$$\left.\frac{\mathrm{d}u}{\mathrm{d}v}\right|_{x=c} = \left.\frac{\Delta u}{\Delta v}\right|_{\substack{x=a,\\ \Delta x=b-a}};$$

the difference quotient is equal to the derivative somewhere in between. Lagrange's theorem is the special case of Cauchy's theorem where $g(x)$ always equals x (so $v = x$), and Rolle's theorem is the special case of Lagrange's theorem where $f(a)$ equals $f(b)$ (so $\Delta u = 0$). By default, *the* Mean-Value Theorem (MVT) is Lagrange's version.

Another useful theorem is **Darboux's Theorem**, a form of the Intermediate Value Theorem (IVT) for derivatives. The usual IVT (on page 18) applies to continuous functions. Now, the derivative of a differentiable function might not be continuous (if it is, then the original function is *continuously* differentiable). However, the derivative of any differentiable function still satisfies the IVT! To be explicit:

> If a function f is defined on at least an interval $[a, b]$ (with $a < b$) such that f is differentiable on all of $[a, b]$, and if L is between $f'(a)$ and $f'(b)$, then $L = f'(c)$ for some c between a and b.

In other words, you can always solve an equation giving the value of a derivative, as long as you already have values of the derivative on either side of the desired value (assuming that the function is differentiable everywhere in between). The standard example of a differentiable function that is not continuously differentiable is $f(x) = x^2 \sin(1/x)$ with $f(0) = 0$; then $f'(x) = 2x \sin(1/x) - \cos(1/x)$ with $f'(0) = 0$. This is discontinuous at 0, but only because f' takes *all* values between -1 and 1 infinitely often approaching 0 (from either side). The value which *should* be the limit, $f'(0) = 0$, is *not* the limit (so f' is not continuous there); however, it still lies within the range $[-1, 1]$ of values (the cluster values) that the function is approaching.

For another variation on the theme of the IVT, suppose that f is continuous on $[a, b]$ and differentiable on (a, b) with f' never zero on (a, b). Then if L is between $f(a)$ and $f(b)$, there is a *unique* number c between a and b such that $L = f(c)$. The IVT tells us that there is *at least* one solution, but Rolle's Theorem says that, if there were two, then f' would have to be zero between them, which it is not. Therefore, there is one and *only* one solution.

4.5 *L'Hôpital's Rule*

One important consequence of Cauchy's mean-value theorem is **L'Hôpital's Rule**. This is a rule for limits again, but it handles limits with forms such as $\infty \div \infty$ and $0 \div 0$.

L'Hôpital's Rule applies when taking limits in any direction D, if u and v are two quantities defined in the direction D, so long as either $\lim_D (1/v) = 0$ (so in other words, $\lim_D v = \pm\infty$) or both $\lim_D u$ and $\lim_D v$ are zero. In that case, if $\lim_D (\mathrm{d}u/\mathrm{d}v)$ exists, then $\lim_D (u/v)$ also exists and the two limits are equal.

L'Hôpital's Rule can also be applied to limits with exponents by taking logarithms, applying the rule directly, and reversing the logarithms. It is therefore very versatile, although Taylor series (see page 61) can do even more.

4.6 *Concavity*

There are various terms used when the values of a function, its average rates of change, or its second average rates of change (the rates of change of the rates of change) are all positive (or negative), at least on some interval. When the function is differentiable, and especially when it's twice differentiable, there are easier ways to describe these. This is all summarized in the table below.

Property of f:	Definition:	If differentiable:	If twice differentiable:
Positive	$f(a) > 0$	—	—
Negative	$f(a) < 0$	—	—
Increasing	$\frac{f(b) - f(a)}{b - a} > 0$	$f'(a) > 0$	—
Decreasing	$\frac{f(b) - f(a)}{b - a} < 0$	$f'(a) < 0$	—
Concave upward	$\frac{\frac{f(c) - f(b)}{c - b} - \frac{f(b) - f(a)}{b - a}}{c - a} > 0$	$\frac{f'(b) - f'(a)}{b - a} > 0$	$f''(a) > 0$
Concave downward	$\frac{\frac{f(c) - f(b)}{c - b} - \frac{f(b) - f(a)}{b - a}}{c - a} < 0$	$\frac{f'(b) - f'(a)}{b - a} < 0$	$f''(a) < 0$

In all of these, the function f has the given property on some interval if the given condition holds whenever a, b, and c are *distinct* numbers in that interval. (They must be distinct to avoid division by zero.)

Generally, it's much easier to work with the rightmost condition for every property, but you can't do that if the necessary derivatives don't exist. Even if the function isn't differentiable at all, it still makes sense to say whether or not it's concave upward or downward.

Incidentally, here is some other terminology that you may see for these properties:

- Sometimes people use $\geq$ and $\leq$ in place of $>$ and $<$. If you want to be clear, you can use adverbs: 'strictly' for the definitions above (using $>$ and $<$) or 'weakly' for the versions with $\geq$ and $\leq$.
- Sometimes people put the word 'monotone' in front of 'increasing' and 'decreasing', even though it really isn't necessary. (However, when people use this word, they are more likely to mean ‹weakly› too, even if they don't say so.)
- Alternatively, if the word 'monotone' is used alone, then it means ‹increasing›; the corresponding word for ‹decreasing› is 'antitone' (but this word is fairly rare). (Again, people who use this terminology are more likely to mean ‹weakly›.)
- If the word 'concave' is used alone, then it means ‹concave downard›; the corresponding word for ‹concave upward› is 'convex' (and this word is extremely common). (People who use this terminology are also more likely to mean ‹weakly›.)

4.7 *Graphing*

If you want to have a complete graph of $y = f(x)$, then these are all of the things that you should make sure show up:

- $x = 0$, if $f(0)$ exists;
- $x \to \infty$, if $f(x)$ exists in that direction;
- $x \to -\infty$, if $f(x)$ exists in that direction;
- $x \to c^-$, if $f(x)$ exists in that direction, whenever f is undefined or discontinuous at c;
- $x \to c^+$, if $f(x)$ exists in that direction, whenever f is undefined or discontinuous at c;
- $x = c$, if $f(c)$ exists, whenever f is undefined approaching c from either direction (or both);
- $x = c$, whenever $f(c) = 0$;
- $x = c$, if $f(c)$ exists, whenever f' is undefined or discontinuous at c;
- $x = c$, whenever $f'(c) = 0$;
- $x = c$, if $f(c)$ exists, whenever f'' is undefined or discontinuous at c;
- $x = c$, whenever $f''(c) = 0$.

This should be sufficient whenever f is a twice-differentiable function whose domain is an interval, or more generally whenever f is *piecewise twice-differentiable*: a piecewise-defined function in which the domain of each piece is an interval and in which each piece is twice-differentiable except possibly at its endpoints. (There are weirder functions that can't be put in this form, but you shouldn't have to deal with them in this class.)

If you have a graphing calculator, then you may use it, but you still need to ensure that all of the features listed above appear. At the very least, this may require you to adjust the calculator's graphing window. If you're graphing by hand, then you'll get the best results if you know the values or limits of f, f', and f'' for all of these places or limits, but you should at least get f for all of them and f' wherever you looked because of something involving f' or f''. You can also look at points in between these (if f is defined there) for an even more precise graph.

There is one other feature of a graph of a function that is easy to see using its derivative: its linear asymptotes. Roughly speaking, two curves are *asymptotes* in a particular direction if they get arbitrarily close together (without meeting) in that direction. By 'direction' here, I mean a direction in the plane where you're graphing, rather than a direction along the x-axis as you're used to from limits; that's just one thing that makes a precise general definition of 'asymptote' beyond the scope of this course. (Another is how to precisely define what a curve is, although you'll see that if you take Calculus 2.) However, when one curve is the graph of a function and the other is a straight line, then we can define these precisely:

The graph of $y = f(x)$ has a **vertical asymptote** $x = c$ in the upward direction whenever $f(c^+)$ or $f(c^-)$ (or both) is ∞, and in the downward direction whenever $f(c^+)$ or $f(c^-)$ (or both) is $-\infty$. That is, we're considering limits of the form $\lim_{x \to c^+} f(x)$ and $\lim_{x \to c^-} f(x)$ and seeing if any of them is $\pm\infty$. If you follow the steps for graphing on the previous page, then you've already found when this happens, so just draw a dashed line for the vertical asymptote. Besides that, the graph has a non-vertical **linear asymptote** $y = mx + b$ in the rightward direction whenever $\lim_{x \to \infty} \big(f(x) - (mx + b)\big) = 0$, and in the leftward direction whenever $\lim_{x \to -\infty} \big(f(x) - (mx + b)\big) = 0$. These take a little more work; if you found that $f(\infty)$ or $f(-\infty)$ (or both) is b, then the horizontal line $y = b$ is an asymptote. But otherwise, if $f'(\infty)$ or $f'(-\infty)$ exists, then you can still find an oblique (also called slant or diagonal) asymptote. First, the limit of f' is the slope m of the asymptote; then you take the limit of $f(x) - mx$ as $x \to \pm\infty$ to find b.

Here is a summary of the rules for asymptotes:

- If $f(x) \to \infty$ as $x \to c^+$, then $x = c$ is a vertical asymptote in the upward direction;
- If $f(x) \to \infty$ as $x \to c^-$, then $x = c$ is again a vertical asymptote in the upward direction;
- If $f(x) \to -\infty$ as $x \to c^+$, then $x = c$ is a vertical asymptote in the downward direction;
- If $f(x) \to -\infty$ as $x \to c^-$, then $x = c$ is again a vertical asymptote in the downward direction;
- If $f(x) \to b$ as $x \to \infty$, then $y = b$ is a horizontal asymptote in the rightward direction;
- If $f(x) \to b$ as $x \to -\infty$, then $y = b$ is a horizontal asymptote in the leftward direction;
- If $f'(x) \to m$ as $x \to \infty$ and $f(x) - mx \to b$ as $x \to \infty$, then $y = mx + b$ is an oblique asymptote in the rightward direction;
- If $f'(x) \to m$ as $x \to -\infty$ and $f(x) - mx \to b$ as $x \to -\infty$, then $y = mx + b$ is an oblique asymptote in the leftward direction.

For functions that aren't differentiable, an oblique asymptote might exist without being found this way, but you should only need to find asymptotes when the function is sufficiently differentiable for this method to work.

4.8 *Optimization*

Literally, **optimization** is making something the best, but we use it in math to mean **maximization**, which is making something the biggest. (You can imagine that the thing that you're maximizing is a numerical measure of how good the thing that you're optimizing is.) Essentially the same principles apply to **minimization**, which is making something the smallest. (And *pessimization* is making something the worst, although people don't use that term very much.) A generic term for making something the largest or smallest is **extremization**.

In theory, optimization is simply finding absolute extrema, which is most easily done for continuous functions on compact intervals. In that case, the maximum and minimum must both exist, by the Extreme Value Theorem (page 19), and each of them must occur at either the endpoint of the interval or where the derivative of the function is either zero or undefined (by the first theorems cited in Section 4.4). However, practical problems cannot always be modelled in this way, so we will need some more general techniques.

The key principle of applied optimization is this:

> A quantity u can only take a maximum or minimum value when its differential $\mathrm{d}u$ is zero or undefined.

If you write u as $f(x)$, where f is a fixed differentiable function and x is a quantity whose range of possible values you already understand (typically an interval), then $\mathrm{d}u = f'(x)\,\mathrm{d}x$. So u can only take an extreme value when its derivative (with respect to x) is zero or undefined or when you can no longer vary x however you please (which must occur at the extreme values of x and typically only then). This recreates the situation that I referred to above, finding the extreme values of a function defined on an interval. However, the principle that $\mathrm{d}u$ is zero or undefined applies even when u is not explicitly given as a function of anything else.

Be careful, because u might not have a maximum or minimum value! Assuming that u varies continuously (which it must if Calculus is to be useful at all), then it must have a maximum and minimum value whenever the range of possibilities is *compact*; this means that if you pass continuously through the possibilities in any way, then you are always approaching some limiting possibility. (In terms of $u = f(x)$, this is the case when f is continuous and its domain, the range of possible values of x, is a compact interval.)

However, if the range of possibilities heads off to infinity in some way, or if there is an edge case that's not quite possible to reach, then you also have to take a limit to see what value u is approaching. (In terms of $u = f(x)$, if the interval is open or unbounded at either end, then there is a direction in which x could vary but in which there is no limiting value of x in the range of possibilities.) If any such limit is larger than every value that u actually reaches (which includes the possibility that a limit is ∞), then u has no maximum value; if any such limit is smaller than every value that u actually reaches (which includes the possibility that a limit is $-\infty$), then u has no minimum value.

So in the end, you look at these possibilities:

- when the derivative of u is zero or undefined,
- the extreme edge cases, and
- the limits approaching impossible limiting cases.

The largest value of u that you find in this way (regardless of whether this value is actually attained or is only approached in the limit) is called the *supremum* of u; similarly, the smallest value of u that you find is called the *infimum* of u. If u actually takes the value of its supremum, then that same value is also the *maximum* of u; but if u only approaches its supremum in a limit, then it has no maximum. Similarly, if u actually takes the value of its infimum, then that same value is also the *minimum* of u; but if u only approaches its infimum in a limit, then it has no minimum.

4.9 *Economic applications*

In word problems in economics or finance, a few quantities arise regularly, which you should know about.

- **Quantity** in this context has a specific meaning: the amount of a good or service made and/or sold in a given period of time. Quantity is thus measured in such units as pounds per week, items per year, or litres per hour. Quantity is variously denoted q or x.
- **Price** (or *unit price*) is the amount of money received for a given amount of goods or services. So price is measured in units such as dollars per pound or euros per item. Price is denoted p, a *lowercase* letter.
- **Revenue** is the amount of money received for goods or services in a given period of time. Revenue is measured in dollars per week, euros per year, etc. Revenue is denoted R, and we have this equation:

  $$R = qp.$$

  (Notice that the units make sense in this equation; amount over time, multiplied by money over amount, becomes money over time.)
- **Cost** is the amount of money that the business has to spend (in a given period of time) in order to produce and distribute their goods and services. (In this terminology, *cost* is completely different from *price*.) Like revenue, cost is measured in units of money over time.
- Finally, **profit** is the amount of money that the business makes and keeps in a given period of time. Unlike everything else here, it makes sense for profit to be negative. Profit is denoted P, an *uppercase* letter, and we have another equation:

  $$P = R - C.$$

In business, you generally want to maximize profit: make it not only positive but as large as possible. Even if you don't want to maximize profit as normally measured (because you care about something else besides money), economists typically try to calculate whatever else you care about and still say that you maximize profit (in a generalized sense).

For any of these quantities, we can discuss their average or marginal values. In this context, the **average** profit/cost/etc is the profit/cost/etc divided by the quantity:

$$\bar{P} = \frac{P}{q},\ \bar{C} = \frac{C}{q},\ \ldots.$$

(As you can see, a bar is used to indicate this ratio. Be careful; when we get to applications of integrals, this bar will be used to denote an average in a different way.) On the other hand, the **marginal** profit/cost/etc is the derivative of profit/cost/etc with respect to quantity:

$$P' = \frac{\mathrm{d}P}{\mathrm{d}q},\ C' = \frac{\mathrm{d}C}{\mathrm{d}q},\ \ldots.$$

(As you can see, a prime tick is used to indicate this derivative, which is safe in context because it always means the derivative respect to q. For a derivative with respect to time, which is also important in this context even though we aren't doing any examples of that in this class, a dot may be used instead.) Although the units for a marginal or average quantity are the same, they represent different things!

Finally, people also speak of the **marginal average** profit/cost/etc:

$$\begin{aligned} \bar{P}' &= \frac{\mathrm{d}(P/q)}{\mathrm{d}q} = \frac{qP' - P}{q^2} = \overline{P' - \bar{P}}, \\ \bar{C}' &= \frac{\mathrm{d}(C/q)}{\mathrm{d}q} = \frac{qC' - C}{q^2} = \overline{C' - \bar{C}}, \\ &\vdots \end{aligned}$$

The marginal profit is particularly important, since it must be zero when profit is maximized (as long as the maximum profit occurs when it is still possible to vary the quantity in any way desired); and since the marginal marginal profit (the second derivative of profit with respect to quantity) is typically negative, the profit really will be maximized when the marginal profit is zero. However, in the absence of information about the revenue, there is a rule of thumb that one should minimize the average cost instead, which means finding where the marginal average cost is zero.

5 Integrals

Besides differentiation, the other major topic of Calculus is integration.

5.1 *Definite integrals*

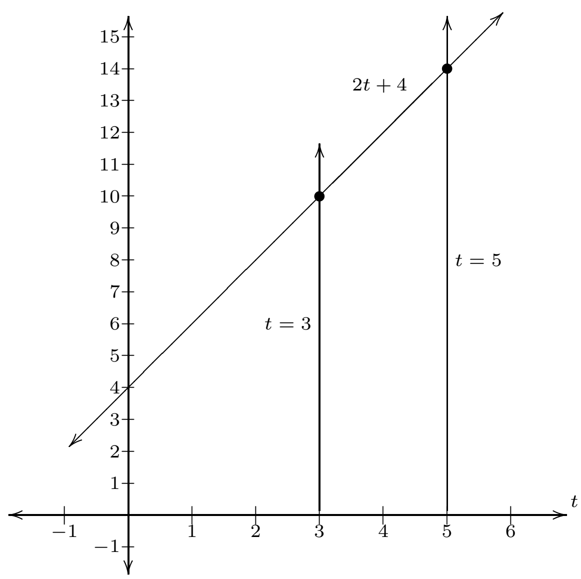
Just as the *differential* of a standard-sized quantity (neither infinite nor infnninitesimal) is an infinitesimal (infinitely small) change in that quantity, so the **definite integral** of an infinitesimal quantity is the sum of infinitely many values of that quantity, giving (typically) a standard-sized result. If x and y are standard quantities, then $y\,\mathrm{d}x$ is a typical infinitesimal quantity. (An expression like this is called a *differential form.*) If we add this up from the point where $x = a$ to the point where $x = b$, then we get the **definite integral**

$$\int_{x=a}^{b} y\,\mathrm{d}x.$$

As long as the same variable x is used throughout, then it's safe to abbreviate this as

$$\int_{a}^{b} y\,\mathrm{d}x.$$

For example, $\int_3^5 (2t+4)\,\mathrm{d}t$ is the sum, as t varies smoothly from 3 to 5, of the product of $2t+4$ and $\mathrm{d}t$ (the infinitesimal change in t) at each stage along the way. We can think of this product as giving the area of a rectangle whose height is $2t+4$ and whose width is $\mathrm{d}t$; if we line these rectangles up side by side, then they combine to give a trapezoid:



We can find out the area of this trapezoid using geometry, since its width is $5 - 3 = 2$ and its height varies linearly from $2(3) + 4 = 10$ to $2(5) + 4 = 14$. Therefore,

$$\int_3^5 (2t+4)\,\mathrm{d}t = \frac{10+14}{2} \cdot 2 = 24.$$

Normally, you can't evaluate an integral by drawing a picture like this; I'll come back to how we can calculate it after a brief digression.

5.2 *Antidifferentials*

If $\mathrm{d}u = y\,\mathrm{d}x$, then $y\,\mathrm{d}x$ is the *differential* of u, as you know. We also say that u is an **antidifferential** of $y\,\mathrm{d}x$. However, u is *not the only* antidifferential of $y\,\mathrm{d}x$; if C is any constant, then $\mathrm{d}(u + C) = y\,\mathrm{d}x$ too, so $u + C$ is also an antidifferential of $y\,\mathrm{d}x$. However, for a continuously defined quantity, there is *no other* antidifferential of $y\,\mathrm{d}x$. Even if there are gaps in the definition of the quantity, we can say that $u + C$ is an antidifferential of $\mathrm{d}u$ if and only if C is a *local* constant, meaning that it can change value only across a gap where u is undefined. (Ultimately, this is a consequence of the theorem that if the derivative of a function on an interval is always zero, then that function must be a constant; the relevant function here is the difference between the functions that give any two possible antidifferentials.)

Antidifferentials are denoted by '$\int$', so we have

$$\int \mathrm{d}u = u + C$$

by definition. (This looks similar to the notation for a definite integral, which makes sense for reasons that will be explained below, but you can tell the difference because there are no bounds attached to the symbol.) For example,

$$\mathrm{d}(t^2 + 4t) = 2t\,\mathrm{d}t + 4\,\mathrm{d}t = (2t + 4)\,\mathrm{d}t,$$

so

$$\int (2t + 4)\,\mathrm{d}t = \int \mathrm{d}(t^2 + 4t) = t^2 + 4t + C.$$

As $2t + 4$ is the derivative of $t^2 + 4t$ with respect to t, we also say that $t^2 + 4t$ is an **antiderivative** of $2t + 4$ with respect to t. An antidifferential or antiderivative is also called an **indefinite integral**; so 'indefinite integral of $(2t + 4)\,\mathrm{d}t$' (antidifferential) and 'indefinite integral of $2t + 4$ with respect to t' (antiderivative) both mean $\int (2t + 4)\,\mathrm{d}t = t^2 + 4t + C$.

To find antidifferentials (or antiderivatives), we must run the rules for differentials (and derivatives) backwards. In the table below, I have some rules for differentiation (all of which you should know by now), together with corresponding rules for integration:

$$\begin{aligned}
\mathrm{d}(u + v) &= \mathrm{d}u + \mathrm{d}v, & \int (y + z)\,\mathrm{d}x &= \int y\,\mathrm{d}x + \int z\,\mathrm{d}x;\\
\mathrm{d}(ku) &= k\,\mathrm{d}u \text{ (when } k \text{ is constant)}, & \int ky\,\mathrm{d}x &= k\int y\,\mathrm{d}x \text{ (when } k \text{ is constant)};\\
\mathrm{d}(uv) &= v\,\mathrm{d}u + u\,\mathrm{d}v, & \int u\,\mathrm{d}v &= uv - \int v\,\mathrm{d}u;\\
\mathrm{d}(u^n) &= nu^{n-1}\,\mathrm{d}u \text{ (when } n \text{ is constant)}, & \int u^m\,\mathrm{d}u &= \frac{1}{m+1}u^{m+1} + C \text{ (when } m \neq -1 \text{ is constant)};\\
\mathrm{d}(\mathrm{e}^u) &= \mathrm{e}^u\,\mathrm{d}u, & \int \mathrm{e}^u\,\mathrm{d}u &= \mathrm{e}^u + C;\\
\mathrm{d}(\ln|u|) &= \frac{1}{u}\,\mathrm{d}u, & \int \frac{1}{u}\,\mathrm{d}u &= \ln|u| + C;\\
\mathrm{d}(\sin u) &= \cos u\,\mathrm{d}u, & \int \cos u\,\mathrm{d}u &= \sin u + C;\\
\mathrm{d}(\cos u) &= -\sin u\,\mathrm{d}u, & \int \sin u\,\mathrm{d}u &= -\cos u + C;\\
&\text{etc.}
\end{aligned}$$

Using these rules, you can work out all of the integrals in the textbook through Chapter 6, and then some.

For example, to find $\int (2t + 4)\,\mathrm{d}t$:

$$\int (2t + 4)\,\mathrm{d}t = \int 2t\,\mathrm{d}t + \int 4\,\mathrm{d}t = 2\int t^1\,\mathrm{d}t + 4\int \mathrm{d}t = 2\left(\frac{1}{2}t^2\right) + 4t + C = t^2 + 4t + C.$$

This is the same answer as earlier on this page, but this time I didn't have to guess the answer and get lucky; I was able to actually calculate it. That's how you're going to be doing most of the problems.

For more complicated integrals, there are fancier techniques. Rather than learn all of these, you can program them into a computer. There are even free websites that will do this for you! For a quick job, you can usually get an answer from Wolfram|Alpha (`http://wolfram.alpha.com/`), although some subtler problems work better using Maxima, which is available online as part of the SageMath system at CoCalc (`https://cocalc.com/`), which takes a little more set-up.

5.3 *The Fundamental Theorem of Calculus*

The **Fundamental Theorem of Calculus** relates definite and indefinite integrals. There are two parts:

1. $\mathrm{d}\Big(\int_{t=a}^{b} f(t)\,\mathrm{d}t\Big) = f(b)\,\mathrm{d}b - f(a)\,\mathrm{d}a$;
2. $\int_{t=a}^{b} \mathrm{d}\big(f(t)\big) = f(b) - f(a)$.

The first part applies whenever f is a fixed continuous function (assuming that a and b are differentiable quantities); in particular, it claims that the integral exists and is differentiable. The second part applies whenever f is a fixed differentiable function (assuming that t is a differentiable quantity); in particular, it claims that the integral exists.

Although both of these parts refer directly to definite integrals, indefinite integrals (antidifferentials) appear implicitly because of the presence of the differentials. Specifically, the first part claims that the definite integral that appears in it is an antidifferential of the differential form on its right-hand side, and the second part shows how to evaluate a definite integral of a differential form whose antidifferential is known.

If you want to express these without refering to the function f, then you can write them thus:

1. $\mathrm{d}\Big(\int_{x=a}^{b} \omega\Big) = \omega|_{\substack{x=b,\\ \mathrm{d}x=\mathrm{d}b}} - \omega|_{\substack{x=a,\\ \mathrm{d}x=\mathrm{d}a}}$;
2. $\int_{x=a}^{b} \mathrm{d}u = u|_{x=b} - u|_{x=a}$.

Here, I'm using ω to stand for an entire differential form (for which people often use Greek letters). These basically say that d and $\int$ cancel as long as you move the bounds on the integral into bounds on a difference (although you also have to do this inside the differentials for the first part). It's very common to introduce the abbreviation $u|_a^b$ for $u|_b - u|_a$; then the second part can also be written thus:

2. $\int_{x=a}^{b} \mathrm{d}u = u|_{x=a}^{b}$.

(This could also be written as $\Delta u|_{\substack{x=a,\\ \Delta x=b-a}}$; this is usually not particularly convenient, although it has some theoretical uses.)

It's the second part of the theorem that we use the most. If you want to evaluate a definite integral $\int_a^b y\,\mathrm{d}x$, then you should first figure out the indefinite integral $\int y\,\mathrm{d}x$. If the answer to this is u (or more generally $u + C$), then this means that $y\,\mathrm{d}x = \mathrm{d}u$; that is, u is an antidifferential of $y\,\mathrm{d}x$. Therefore, $\int_{x=a}^{b} y\,\mathrm{d}x = \int_{x=a}^{b} \mathrm{d}u$, and the FTC tells us that this is equal to $u|_{x=a}^{b}$. As this last expression is simply a difference, you can figure it out using simple algebra.

For example, consider

$$\int_{t=3}^{5} (2t+4)\,\mathrm{d}t.$$

In the last section, we saw that $\int (2t+4)\,\mathrm{d}t = t^2 + 4t + C$; in other words, $(2t+4)\,\mathrm{d}t = \mathrm{d}(t^2+4t)$. Therefore,

$$\begin{aligned}\int_3^5 (2t+4)\,\mathrm{d}t &= \int_3^5 \mathrm{d}(t^2+4t) = (t^2+4t)|_3^5 \\ &= \big((5)^2 + 4(5)\big) - \big((3)^2 + 4(3)\big) = (45) - (21) = 24.\end{aligned}$$

(Notice that this is the same answer as when I did this using geometry on page 41.)

This also explains why the same term 'integral' and symbol '$\int$' are used for both the definite integral (a sum of infinitely small quantities) and the indefinite integral (the antidifferential). They at first appear to be completely different concepts, but in reality they are closely related, through the Fundamental Theorem of Calculus.

So in summary, to find the indefinite integral $\int y \, \mathrm{d}x$, you need to use integration techniques; your answer will still have the variable in it and should end with a new local-constant term C. To find the definite integral $\int_a^b y \, \mathrm{d}x$, first find the indefinite integral and then take a difference; assuming a and b are constants, your answer will also be constant (and the C will disappear).

So for example, to find the definite integral of $2t + 4$ with respect to t from 3 to 5:

$$\int_3^5 (2t + 4) \, \mathrm{d}t = \int_3^5 (2t^1 \, \mathrm{d}t + 4 \, \mathrm{d}t) = \left(2\left(\frac{1}{2} t^2 \right) + 4t \right) \Big|_3^5 = (t^2 + 4t)|_3^5 = 45 - 21 = 24.$$

This is simply a combination of calculations that I did earlier, to find the indefinite integral and to apply the FTC.

5.4 *Semidefinite integrals*

Besides the *definite* integral $\int_a^b f(x) \, \mathrm{d}x$ and the *indefinite* integral $\int f(x) \, \mathrm{d}x$, there is also a **semidefinite integral** $\int_a f(x) \, \mathrm{d}x$. While the definite integral works out to a specific value (as long as f, a, and b are specified), the indefinite and semidefinite integrals still have the variable x in them. On the other hand, while the indefinite integral depends on an arbitrary C, the definite and semidefinite integrals don't have this. So the semidefinite integral fits in between the other two kinds.

Here is one way to define it:

$$\int_{x=a} f(x) \, \mathrm{d}x = \int_{t=a}^{x} f(t) \, \mathrm{d}t.$$

That is, introduce a new variable t and use the old variable x as the upper bound of a definite integal. The Second Fundamental Theorem of Calculus,

$$\int_{x=a}^{b} f(x) \, \mathrm{d}x = \left(\int f(x) \, \mathrm{d}x \right) \Big|_{x=a}^{b} = \left(\int f(x) \, \mathrm{d}x \right) \Big|_{x=b} - \left(\int f(x) \, \mathrm{d}x \right) \Big|_{x=a},$$

also tells us how to evaluate semidefinite integrals:

$$\int_{x=a} f(x) \, \mathrm{d}x = \int f(x) \, \mathrm{d}x - \left(\int f(x) \, \mathrm{d}x \right) \Big|_{x=a}.$$

In other words, work out the indefinite integral as usual; then, instead of evaluating this at two values of the variable before subtracting, evalute it at one value and keep the variable in the other expression (then subtract). For example,

$$\int_{x=1} x \, \mathrm{d}x = \frac{x^2}{2} - \left(\frac{x^2}{2} \right) \Big|_{x=1} = \frac{x^2}{2} - \left(\frac{(1)^2}{2} \right) = \frac{1}{2} x^2 - \frac{1}{2}.$$

(You can probably skip the step with $|_{x=1}$ in it, since once you've written down $x^2/2$ before the minus sign, you can immediately plug in 1 for x to get $(1)^2/2$ after the minus sign.)

5.5 *Integration by parts*

Integration by parts is based on the Product Rule for differentiation. In terms of differentials, the Product Rule says that $\mathrm{d}(uv) = v \, \mathrm{d}u + u \, \mathrm{d}v$. Taking indefinite integrals of both sides and rearranging the terms slightly, this becomes

$$\int u \, \mathrm{d}v = uv - \int v \, \mathrm{d}u.$$

Unlike integration by substitution, you don't rewrite the problem in terms of u (nor v). Instead, you identify suitable u and v and their differentials and then write out the equation above in terms of x (or whatever your variable is).

You want to pick u and v so that $\int u \, \mathrm{d}v$ is the integral that you care about, which means splitting up the factors of the integrand, some into u and some into $\mathrm{d}v$. Once you know u and $\mathrm{d}v$, you can find $\mathrm{d}u$ and v, at least if you know how to integrate whatever $\mathrm{d}v$ is. (When you do this integration of $\mathrm{d}v$ to get

v, you have a choice up to a local constant; you're deciding what v is, so just pick the simplest expression.) If you split things up well, then $\int v \, \mathrm{d}u$ will be simpler than what you started with.

Here is my advice on how to split factors into u and $\mathrm{d}v$ so that integration by parts will make the next integral easier. The items on the top of the list below are the best choices for $\mathrm{d}v$, and the items on the bottom are the best choices for u. Put as many factors as you can into $\mathrm{d}v$, starting at the top of this list and working your way to the bottom, as long as you still have something that you know how to integrate to get v. Then put whatever factors are left over into u.

- $\mathrm{d}x$ (this *must* go into $\mathrm{d}v$),
- e^x and other exponential expressions,
- $\sin x$ and other trigonometric expressions,
- polynomials and other algebraic expressions,
- $\ln x$ and other logarithmic expressions,
- $\operatorname{asin} x = \sin^{-1} x$ and other inverse trigonometric expressions.

In complicated cases, you may have to use integration by parts more than once. Just keep going until either you get something that you can handle or you get back to where you started. In the latter case, you can set up an equation to solve for your integral.

5.6 *Geometric applications*

Given a region in a plane, you could consider its area directly or revolve it around a line and consider the volume of the resulting **solid of revolution**. (Of course, you could do other things, but these are the only things that we're considering in this course. Come back in Calculus 3 if you want more.) For simplicity, suppose that the region has two parallel line segments as opposite sides and every line between those passes through the region in only one segment. Then if you set up a coordinate system so that one axis is parallel to these two parallel sides, then there are numbers a and b (with $a \leq b$) and functions f and g (defined on $[a, b]$, with $f \geq g$), so that the region is given by the inequalities $a \leq x \leq b$ and $f(x) \geq y \geq g(x)$ in the variables x and y. If f and g are continuous, then we can also say that the region is bounded by $x = a$, $x = b$, $y = f(x)$, and $y = g(x)$, and the area of the region is

$$\int_{x=a}^{b} \big(f(x) - g(x)\big) \, \mathrm{d}x.$$

If a region doesn't have the appropriate shape, it may still be possible to divide it into regions with such a shape. You can also swap x and y if that makes the region easier to describe or the integral easier to compute.

If the region is revolved around a line in the plane, then this is simple to describe only when this **line of revolution** is parallel or perpendicular to the region's parallel boundary lines, in which case you can set up the coordinate system so that the line of revolution is one of the coordinate axes. If you revolve the region described above around the x-axis, then the volume of the resulting solid of revolution is

$$\int_{x=a}^{b} \pi \big(f(x)^2 - g(x)^2\big) \, \mathrm{d}x,$$

assuming that f and g are continuous and $g \geq 0$. If instead you revolve this region around the y-axis, then the volume of the resulting solid of revolution is

$$\int_{x=a}^{b} 2\pi x \big(f(x) - g(x)\big) \, \mathrm{d}x,$$

assuming that f and g are continuous and $a \geq 0$. If the line of revolution is *not* parallel or perpendicular to parallel boundary lines, then you need to describe the region in a more complicated way by dividing it into regions with boundary lines that *are* parallel or perpendicular to the line of revolution. However, you will learn how to find the volumes of much more general solids if you take Calculus 3.

If instead of a region in the plane, you start with a *curve* in the plane, then we can only handle this for now if the curve is a graph of a differentiable function f. Specifically, if $a \leq b$ and f is defined on $[a, b]$, then the length of the graph of $y = f(x)$ is

$$\int_{x=a}^{b} \sqrt{f'(x)^2 + 1} \, \mathrm{d}x,$$

assuming that f' is continuous on $[a, b]$. You'll learn how to handle more general curves if you take Calculus 2 (at the very end) and even more in Calculus 3 (towards the beginning).

If instead you revolve this graph around the x-axis, then the area of the resulting **surface of revolution** is

$$\int_{x=a}^{b} 2\pi f(x) \sqrt{f'(x)^2 + 1} \, \mathrm{d}x,$$

assuming that f' is continuous on $[a, b]$ and $f \geq 0$. Finally, if you revolve this same curve around the y-axis, then the area of the resulting surface is

$$\int_{x=a}^{b} 2\pi x \sqrt{f'(x)^2 + 1} \, \mathrm{d}x,$$

assuming that f' is continuous on $[a, b]$ and $a \geq 0$. (For some reason, this formula is not in the textbook.) You will learn how to find the areas of much more general surfaces if you take Calculus 3 (towards the end).

6 Differential equations

A **differential equation** is an equation with differentials or derivatives in it. Here are three examples of differential equations:

$$f'(x) = 3f(x);$$
$$\frac{\mathrm{d}y}{\mathrm{d}x} = 3y;$$
$$\mathrm{d}y = 3y\,\mathrm{d}x.$$

In fact, these three examples are all basically equivalent. If you are given the first of these, then you should make up a name for $f(x)$, say y, and turn the first equation into the middle one. And in the middle equation, you should clear fractions to turn it into the last one. (But any of these might be the original form, depending on how the equation is thought up in the first place.)

6.1 *Separation of variables*

To actually solve this equation, you can use the technique of **separation of variables**. After reaching the last equation, notice that x only appears on the right-hand side but y appears on both sides. If you divide both sides by y, however, then y appears only on the left-hand side. (If $y = 0$, then dividing by y is invalid; I'll come back to that later.) Then the variables are separated:

$$\frac{\mathrm{d}y}{y} = 3\,\mathrm{d}x.$$

(If you're ever unsure which side to put which variable on, then try to put the differentials in the numerators of any fractions. In this example, $1/\mathrm{d}x = 3y/\mathrm{d}y$ would have the variables separated just as much, but it would be less useful, because the next step, below, wouldn't work.)

Now take the indefinite integral of each side of the equation:

$$\int \frac{\mathrm{d}y}{y} = \int 3\,\mathrm{d}x;$$
$$\ln|y| + C_1 = 3x + C_2;$$
$$\ln|y| = 3x + C_2 - C_1.$$

Each integral gives an arbitrary constant, and I subtracted to put them both on the right-hand side. However, since $C_2 - C_1$ could itself be any constant, you can just write this as

$$\ln|y| = 3x + C.$$

In practice, you can skip the other steps with constants and just remember to tack a constant onto the last integral in the equation.

We're not finished; this equation is no longer a differential equation, but it also hasn't been solved for anything. If we want to solve it for y, then we still need to do some algebra to get y by itself on its side of the equation:

$$|y| = \mathrm{e}^{3x+C};$$
$$y = \pm\mathrm{e}^{3x+C}.$$

(If you're given an equation in x and y, then it's a good bet that they want you to solve for y; if you're given an equation like the first example with a function in it, then it's a good bet that they want you to solve for the function. But in principle, you could solve any of these equations for x instead.)

There is one mistake here, which is the step where I divided by y. If $y = 0$, then this is invalid. Furthermore, if $y = 0$ always, then the equation is true, because then both sides of the original equation (in any of the three forms) are 0. (This sort of special exception is fairly common with differential equations.) So a complete solution is

$$y = \pm\mathrm{e}^{3x+C} \text{ or } y = 0.$$

You can make the final solution look a bit nicer by writing $\pm e^{3x+C}$ as $\pm e^{C}e^{3x}$ and then making up a name for $\pm e^{C}$, say P. Since e^{C} could be any positive number, P could be any positive or negative number; the exception $y = 0$ is captured by $P = 0$. So the nicest form of the final solution is

$$y = Pe^{3x},$$

where P is an arbitrary constant. (However, you shouldn't always expect to be able to do a simplifying trick like that.)

Of course, if the original form of the equation is the first example, then you should write this solution as

$$f(x) = Pe^{3x}.$$

6.2 *Initial-value problems*

An **initial-value problem** consists of a differential equation together with enough data to determine the arbitrary constants. Here are three examples of initial-value problems:

$$f'(x) = 3f(x),\ f(0) = 5;$$
$$\frac{dy}{dx} = 3y,\ y|_{x=0} = 5;$$
$$dy = 3y\,dx,\ y|_{x=0} = 5.$$

Again, these three examples are all basically equivalent; if $y = f(x)$, then $y|_{x=0}$ means $f(0)$.

There are two ways to solve an initial-value problem. One is to ignore the initial value and just solve the differential equation, at first. In this example, that gives us

$$y = Pe^{3x},$$

as you've seen. Then you put in the given values, which in this case gives

$$5 = Pe^{3(0)}.$$

Now you can solve for P:

$$5 = P(1);$$
$$P = 5.$$

Therefore, the final answer to the initial-value problem is

$$y = 5e^{3x}.$$

(Again, if the original form of the equation is the first example, then you should write this solution as $f(x) = 5e^{3x}$.)

Another technique is to solve the entire problem at once with the help of *semidefinite integrals* (Section 5.4 on page 44). Let's solve the example

$$dy = 3y\,dx,\ \ y|_{x=0} = 5$$

using semidefinite integrals. Again, separate the variables:

$$\frac{dy}{y} = 3\,dx.$$

Now instead of taking *indefinite* integrals of both sides, take *semidefinite* integrals, using the initial value to guarantee that you're doing the same thing to each side even though it's being done using different variables. In this case, since $y = 5$ when $x = 0$, a semidefinite integral starting at $y = 5$ is the same operation as a semidefinite integral starting at $x = 0$, so

$$\int_{y=5} \frac{\mathrm{d}y}{y} = \int_{x=0} 3 \, \mathrm{d}x.$$

Evaluating these using the FTC gives

$$\ln|y| - \ln|5| = 3x - 3(0).$$

So compared to the integration without the initial value, the difference is that we know which specific constants to use in each integral. Now again, solve for y to finish:

$$\begin{aligned} \ln|y| &= 3x - 0 + \ln 5; \\ |y| &= \mathrm{e}^{3x+\ln 5}; \\ y &= \pm 5\mathrm{e}^{3x}. \end{aligned}$$

This is not completely perfect, because of the $\pm$, but we can figure this out by checking whether y really is 5 when $x = 0$; this will only be true if the sign is $+$. Finally, since we did again divide by y while solving this, check to make sure that y is never zero in the solution; it's not, so the final answer is

$$y = 5\mathrm{e}^{3x}.$$

Of course, this is the same solution as I got before, but this time I got the entire solution all at once without having to first get a solution with an arbitrary constant and then solving for the constant. You may solve intial-value problems using whichever method you prefer.

6.3 *Integrals as solutions to equations*

Although we normally solve a differential equation by taking integrals, you can also think of an integral as a solution to a differential equation. For example, the indefinite integral $\int f(x)\,\mathrm{d}x$ is the solution to the differential equation $\mathrm{d}y/\mathrm{d}x = f(x)$, and the semidefinite integral $\int_{x=a} f(x)\,\mathrm{d}x$ is the solution to the initial-value problem $(\mathrm{d}y/\mathrm{d}x = f(x), y|_{x=a} = 0)$. More generally, the solution to the initial-value problem $(\mathrm{d}y/\mathrm{d}x = f(x), y|_{x=a} = b)$ is $b + \int_{x=a} f(x)\,\mathrm{d}x$. These kinds of initial-value problems are in Sections 4.8 and 5.5 of the textbook and are covered in Calculus 1; more general differential equations and initial-value problems are in Section 7.2 and are covered in Calculus 2.

(There are even more general differential equations than I have discussed here, ones in which it is impossible to separate the variables in the equation; some of these are covered in Chapters 16 and 17 of the online-only version of the textbook. Yet more general differential equations are covered in SCC's course on differential equations, which is basically Calculus 4, but using a different textbook dedicated to that subject. Beyond that, there are graduate-level courses that you could take at a university; in fact, the study of differential equations is a major field of active research in mathematics. We are very far from knowing how to solve them all!)

6.4 *Existence of solutions*

An initial-value problem of the form

$$\mathrm{d}y/\mathrm{d}x = F(x),\ y|_{x=a} = b$$

must have a solution if F is a continuous function, because the solution is simply

$$y = b + \int_{x=a} F(x)\,\mathrm{d}x$$

if the integral exists, and this integral always exists if F is continuous.

A more general case is any *explicit* initial-value problem, in which the left-hand side of the differential equation is still $\mathrm{d}y/\mathrm{d}x$ but the right-hand side is a more general expression involving x and y. (We can write

$$\frac{\mathrm{d}y}{\mathrm{d}x} = F(x, y),$$

where F is a *binary function* or *function of two variables*.) Suppose that the initial value is that $y = b$ when $x = a$. If you fix a value of y near b, then you can think of the right-hand side as a function of x and ask whether this function is continuous near a; if you fix a value of x near a, then you can think of the right-hand side as a function of y and ask whether that function is continuous near b. (If they both are, then we say that the binary function F is *separately continuous* near (a, b).) There is a theorem, called the Peano Existence Theorem, that every initial-value problem satisfying these conditions has a solution. (The functions in these conditions don't even have to be continuous everywhere, no matter what value is chosen for the other variable; even if they are continuous only *near* the initial value, and only when a value *near* the initial value is chosen for the other variable, then y is a function of x for some function that is defined *near* a.)

A stronger result, called the Picard–Lindelöf Theorem, guarantees that the solution is *unique* if in addition, when you fix a value of x near a and think of the right-hand side as a function of y, that function is *differentiable* near b.

7 Sequences and series

For the purposes of this class, a **sequence** is a function whose domain consists only of integers. (It's not necessary that all integers belong to the domain, just that nothing else does.) To emphasize that we're considering a sequence, people often write f_n instead of $f(n)$ when f is a sequence (and n is an integer in its domain). In fact, 'f' is not a very common name for a sequence; 'a' and 'x' (or letters near them) are much more common. Similarly, the argument of a sequence is usually denoted by a letter near the middle of the alphabet (usually between 'i' and 'n'), since these letters are often used for integers. (Still, as with any other variable, you can use any letter that you like in principle.) There is also some redundant terminology: instead of speaking of the input (or argument) and output (or value) of a function, we speak of an **index** and **term** of a sequence. For example, if $a_n = (-2)^n$, then the term with index 3 is $a_3 = (-2)^3 = -8$. (Sometimes people say that 8 is the 3rd term, but this doesn't work so well for $n < 1$.)

7.1 *Limits of sequences*

Since Calculus is about continuously varying quantities and a sequence has only discrete values (at most one for each integer), there's not much Calculus to be done with a sequence. Nevertheless, there is some: you can consider the limit of a sqeuence approaching infinity (or negative infinity). That is, while $\lim_{n \to c} a_n$ (for finite c), $\mathrm{d}a_n/\mathrm{d}n$, and $\int a_n \,\mathrm{d}n$ don't make sense, nevertheless $\lim_{n \to \infty} a_n$ and $\lim_{n \to -\infty} a_n$ can make sense. I'll focus on the first of these, which you can call simply the **limit** of the sequence, because many of our sequences will only be defined at natural numbers; however, limits approaching negative infinity really aren't much different.

Sometimes it's convenient to think of a sequence as the restriction to integers of some more general function. For example, if you're working with the sequence $a_n = 3n^2$, then you can think of the function $f(x) = 3x^2$; while f is defined for all real numbers and a is defined only for integers, otherwise they are the same thing. Since $\lim_{x \to \infty} f(x) = \infty$, this tells us that $\lim_{n \to \infty} a_n = \infty$ too. So most of the time, you can work out the limit of a sequence in the same way that you work out any other limit approaching infinity. If $a_n = f(n)$ for n an integer and f has a limit (possibly infinite) approaching infinity, then a has the same limit; this is a theorem. However, it's possible that a has a limit even when f does not, for example if $f(x) = \sin(\pi x)$. This has no limit as $x \to \infty$, since all values between -1 and 1 are taken for arbitrarily large values of x. When n is an integer, however, $\sin(\pi n) = 0$, so the limit of the sequence $a_n = \sin(\pi n)$ (which is really just the sequence $a_n = 0$) is 0.

There are some more systematic ways of turning a sequence into a function that's defined everywhere (or almost everywhere). These involve the floor and ceiling operations: the **floor** $\lfloor x \rfloor$ of a real number x is the largest integer that's not larger than x, and the **ceiling** $\lceil x \rceil$ of x is the smallest integer that's not smaller than x. Ever since you first learnt to round numbers up and down, you've been using these operations, even if you didn't have names for them; for example, $\lfloor 2.37 \rfloor = 2$ (round down to the nearest integer), and $\lceil 2.37 \rceil = 3$ (round up to the nearest integer). Be careful with negative numbers: $\lfloor -2.37 \rfloor = -3$, and $\lceil -2.37 \rceil = -2$. An important inequality about floors and ceilings is

$$\lfloor x \rfloor \leq x \leq \lceil x \rceil.$$

As long as x is itself fractional (that is not an integer), then

$$\lfloor x \rfloor < x < \lceil x \rceil,$$

and in that case you also have

$$\lfloor x \rfloor + 1 = \lceil x \rceil.$$

(But integers are an exception; if x is an integer, then $\lfloor x \rfloor$, x, and $\lceil x \rceil$ are all equal to each other.)

Using these operations, we can convert any sequence into a function defined more generally: if a is a sequence, then we can consider $a_{\lfloor x \rfloor}$ and $a_{\lceil x \rceil}$. If a is defined for all integers, then these will be defined for all real values of x; even if a isn't defined for all integers, still $a_{\lfloor x \rfloor}$ and $a_{\lceil x \rceil}$ will be defined for many more real numbers. And now we have this theorem:

$$\lim_{x\to\infty} a_{\lfloor x \rfloor} = \lim_{n\to\infty} a_n = \lim_{x\to\infty} a_{\lceil x \rceil}.$$

These functions $a_{\lfloor x \rfloor}$ and $a_{\lceil x \rceil}$ are unusual, since they are (for most sequences) discontinuous at every integer, but they can be handy to think about.

You can see a picture of these in Figure 9.12 on page 519 in Section 9.3 of the textbook. (The textbook is using this picture for a different purpose, although it is related.) In this picture, the book begins with a function f and then constructs a sequence a out of it by defining $a_n = f(n)$. Then on the top (Figure 9.12.a), it shows the graph of $y = f(x)$ in blue along with a graph of $y = a_{\lfloor x \rfloor} = f(\lfloor x \rfloor)$ in magenta; while on the bottom (Figure 9.12.b), it shows a graph of $y = f(x)$ in blue again but now with a graph of $y = a_{\lceil x \rceil} = f(\lceil x \rceil)$ in magenta. You'll notice that the sequence and all three of the other functions tend to the same limit (which in this case is 0). Even if the textbook had started with a function f that did not converge to a limit, the sequence and the two functions defined by floor and ceiling would still all converge to the same thing, or else diverge in the same way.

7.2 *Series*

I wrote above that you can't do much Calculus on sequences; in particular, I remarked that the derivative $\mathrm{d}a_n/\mathrm{d}n$ and integral $\int a_n \,\mathrm{d}n$ don't make sense. Ultimately, this is because $\mathrm{d}n$, an infinitesimal (infinitely small) but non-zero change in n, doesn't make sense when n takes only integer values; the smallest possible non-zero change in n is a change by 1, which is not infinitely small.

But there is something *analogous* to derivatives and integrals. The analogue to derivatives is the **difference** $\Delta_n a_n = a_{n+1} - a_n$, which is the difference of a_n with respect to n. (For example, $\Delta_n(3n) = 3(n+1) - 3n = 3$, and $\Delta_m(m^2) = (m+1)^2 - m^2 = 2m + 1$, which means that if $n = m^2$, then $\Delta_m n = 2\sqrt{n} + 1$.) Whereas the derivative is defined as a limit of difference quotients, the difference simply *is* a difference quotient where the change in n is $\Delta_n n = 1$. (Unfortunately, sequences do not have an analogue of the differential that will take care of changing from one variable to another. This is because $\Delta_u n \cdot \Delta_m u$ bears no particular relationship with $\Delta_m n$, even assuming that all of the values of u are integers.)

The analogue to an integral is a **series**, which is the result of adding up some of the terms of a sequence. (This word can be confusing, in two ways. The first is a quirk of grammar: the plural of 'series' is just 'series' again. You can say 'serieses' as the plural, although this is nonstandard, but using 'serie' as the singular is just plain wrong. The other confusing thing is that, in ordinary language, 'sequence' and 'series' mean basically the same thing; but in mathematics, a sequence is the more basic concept, being essentially just a list of numbers or other quantities, while a series is a sum that you build out of a sequence.)

Like differences, a finite series has no Calculus in it; you just add up some numbers. For example,

$$\begin{aligned}\sum_{n=3}^{7} (n^2+1) &= \big((3)^2+1\big) + \big((4)^2+1\big) + \big((5)^2+1\big) + \big((6)^2+1\big) + \big((7)^2+1\big) \\ &= 10 + 17 + 26 + 37 + 50 = 140.\end{aligned}$$

This means the sum of all of the values of $n^2 + 1$ as n runs from 3 to 7, taking only integer values along the way. That is, it's the sum of all of the values of $n^2 + 1$ as n takes the values 3, 4, 5, 6, and 7, which is what I calculated.

Strictly speaking, this is analogous to a proper integral such as $\int_{x=3}^{8} (x^2+1)\,\mathrm{d}x$; notice that the top number has changed from 7 to 8. Actually, this is more than just an analogy: a series *is* an integral, albeit one whose Calculus content is trivial. Specifically,

$$\sum_{n=i}^{j} a_n = \int_{x=i}^{j+1} a_{\lfloor x \rfloor}\,\mathrm{d}x = \int_{x=i-1}^{j} a_{\lceil x \rceil}\,\mathrm{d}x.$$

(So in this example, $\sum_{n=3}^{7}(n^2+1) = \int_{x=3}^{8}\left(\lfloor x\rfloor^2+1\right)\mathrm{d}x$.) Since these are integrals of piecewise-constant functions, working them out is easy and just results in the original sum. So you don't want to evaluate a series by turning it into an integral; still, it can be handy to know that this can be done, because we know a lot of theorems about integrals that now automatically apply to series.

We traditionally speak of a sum from $n=a$ to $n=b$, written $\sum_{n=a}^{n=b}$ or simply $\sum_{n=a}^{b}$, where $b-a$ is a whole number $(0,1,2,\ldots)$; assuming for simplicity that a is an integer (so that b is also), this sum covers every integer n that satisfies the inequality $a \leq n \leq b$, or in other words all of the integers in the interval $[a,b]$. In light of the relationship between series and integrals, it can be better to think of such a sum as running from $n=a$ to $n=b+1$, but with the last item not quite included; that is, the sum covers every integer n that satisfies the inequality $a \leq n < b+1$, or in other words all of the integers in the interval $[a,b+1)$. Of course, from this perspective, it's not the number b that matters but the number $b+1$; if we call this B, then we can write $\sum_{n=a}^{<B}$ for what is normally written as $\sum_{n=a}^{b}$. Note also that it makes perfect sense to have $B=a$ (in other words, $b-a=-1$); then we are adding up no terms, and the sum is 0.

One nice consequence is that the number of terms in the sum is simply $B-a$ rather than $b-a+1$. Perhaps more importantly, we have this theorem:

$$\sum_{A=n}^{<B} + \sum_{B=n}^{<C} = \sum_{A=n}^{<C},$$

which looks nicer than

$$\sum_{n=a}^{b} + \sum_{n=b+1}^{c} = \sum_{n=a}^{c}.$$

The upshot of all of this is that, when you see (for example) a sum as n runs from 2 to 5, you might want to think of it as a sum over $2 \leq n < 6$ instead.

7.3 *Infinite series*

Besides this, we also consider *infinite* series, which are analogous to infinite improper integrals. Just as $\int_{x=a}^{\infty} f(x)\,\mathrm{d}x$ is defined as $\lim_{b\to\infty}\int_{x=a}^{b} f(x)\,\mathrm{d}x$, so an infinite series is defined as a limit of finite series:

$$\sum_{n=i}^{\infty} a_n = \lim_{j\to\infty}\sum_{n=i}^{j} a_n = \lim_{J\to\infty}\sum_{i=n}^{<J} a_n;$$

the finite sum $\sum_{n=i}^{j} a_n$ or $\sum_{i=n}^{<J} a_n$ is called a **partial sum** of the series. (As with infinite integrals, you can also replace i with $-\infty$, but we won't be doing that very often.) Now there is a limit (and hence Calculus) involved even for series. If this limit converges (to a finite real number), then we say that the infinite series **converges** (to that number); otherwise, it **diverges**. Sometimes it's useful to say that it diverges to ∞ or $-\infty$ (if it does), but this still counts as divergence.

You can also write

$$\sum_{n=i}^{\infty} a_n = \int_{x=i}^{\infty} a_{\lfloor x\rfloor}\,\mathrm{d}x;$$

that is, an infinite series isn't merely *analogous* to an infinite improper integral, it actually *is* an infinite improper integral, even if trying to evaluate this integral just turns it back into the series. Again, look at Figure 9.11.a on page 501 of the textbook; this time, ignore the function f and its blue curve, but notice how the area under the magenta staircase (which is the graph of $a_{\lfloor x\rfloor}$, so the area under it is the integral $\int_{x=1}^{\infty} a_{\lfloor x\rfloor}\,\mathrm{d}x$) represents the infinite sum $a_1 + a_2 + \cdots = \sum_{n=1}^{\infty} a_n$.

It's important to distinguish convergence of a series from convergence of its sequence of terms. If we think of the numbers a_0, a_1, a_2, and so on as the terms of a *sequence*, then this sequence converges if its limit $\lim_{n\to\infty} a_n$ exists; but if we think of them as the terms of a *series*, then this series converges if its sum $\sum_{n=0}^{\infty} a_n$ exists, and this is the limit of a different sequence (the sequence of partial sums).

Nevertheless, there is a relationship between a series and its sequence of terms: the series can only converge if the sequence does, and in fact the series can only converge if the sequence of terms converges to zero! This is because the jth term is

$$a_j = \sum_{n=i}^{j} a_n - \sum_{n=i}^{j-1} a_n = \sum_{n=1}^{<j+1} a_n - \sum_{n=i}^{<j} a_n;$$

if the series converges, then

$$\lim_{j \to \infty} a_j = \sum_{n=i}^{\infty} a_n - \sum_{n=i}^{\infty} a_n = 0$$

(since $j - 1 \to \infty$ or $j + 1 \to \infty$ if and only if $j \to \infty$), but if the series doesn't converge, then this argument is invalid and $\lim_{j \to \infty} a_j$ could be anything. Be careful, however, since this argument only goes one way; if the limit of the sequence of terms is zero, then that tells you *nothing* about whether the series converges.

7.4 *The Fundamental Theorem for series*

In the analogy between sequences and functions, where differentiation of functions corresponds to differences of sequences and integrals correspond to series, there is an analogue of the Fundamental Theorem of Calculus. Just as $(\mathrm{d}/\mathrm{d}x)(\int_{t=a}^{x} f(t)\,\mathrm{d}t) = f(x)$ (the first part), so

$$\Delta_n\left(\sum_{m=i}^{n} a_m\right) = a_{n+1}, \text{ and } \Delta_n\left(\sum_{m=i}^{<n} a_m\right) = a_n.$$

And just as $\int_{x=a}^{b}(F'(x)) = F(b) - F(a)$ (the second part), so

$$\sum_{n=i}^{j}(\Delta_n b_n) = b_{j+1} - b_i, \text{ and } \sum_{n=1}^{<j}(\Delta_n b_n) = b_j - b_i.$$

(In each of these, the analogy works better if you think of $\sum_{n=i}^{j-1}$ as $\sum_{i=n}^{<j}$, as described near the top of page 49.)

The sum of a difference is called a **telescoping series**. A telescoping series converges precisely when the original sequence (not the difference) converges:

$$\sum_{n=i}^{\infty}(\Delta_n b_n) = \lim_{j \to \infty} \sum_{n=1}^{<j}(\Delta_n b_n) = \lim_{j \to \infty}(b_j - b_i) = \lim_{j \to \infty} b_j - b_i.$$

This result is so important that I'll repeat it without the difference notation (which is not widely used):

$$\sum_{n=i}^{\infty}(b_{n+1} - b_n) = \lim_{j \to \infty} b_j - b_i.$$

Sometimes people prefer to write this as

$$\sum_{n=i}^{\infty}(b_n - b_{n-1}) = \lim_{j \to \infty} b_j - b_{i-1}.$$

Just as you can get a list of integrals that you can do by finding the derivatives of basic functions, so you can get a list of series that you can do by finding the differences of basic functions. We could do this with polynomials, for example; although it doesn't come out as simply as in the continuous case, you can derive formulas to sum any polynomial sequence. But an even simpler example is an exponential sequence. That is, consider the difference of r^n with respect to n, where r is constant.

$$\Delta_n(r^n) = r^{n+1} - r^n = r^n(r-1).$$

If anything, this is *simpler* than $\mathrm{d}(r^x)/\mathrm{d}x = r^x \ln x$; the natural logarithm has been replaced by a subtraction. Conversely, if you want to sum r^n, you just need to divide by the constant $r-1$. So

$$\sum_{i=n}^{<J} r^n = \frac{r^J - r^i}{r-1},$$

which is more commonly written as

$$\sum_{n=i}^{j} r^n = \frac{r^i - r^{j+1}}{1-r}.$$

Of course, this doesn't work if $r = 1$; for that, $\sum_{n=i}^{<J} 1^n = J - i$, or $\sum_{n=i}^{j} 1^n = j - i + 1$.

A series like this is traditionally called a **geometric series**. The infinite version converges whenever $\lim_{J\to\infty} r^J$ exists (for $r \neq 1$), which happens precisely when $|r| < 1$, in which case the limit is actually 0. (If $r > 1$, then the limit is ∞; if $r = 1$, then the limit is $\lim_{J\to\infty} J = \infty$; if $r = -1$, then it oscillates between 1 and -1; and if $r < -1$, then it oscillates between ∞ and $-\infty$.) Therefore,

$$\sum_{n=i}^{\infty} r^n = -\frac{r^i}{r-1} = \frac{r^i}{1-r}$$

if $|r| < 1$.

7.5 *Convergence tests*

Here is a summary of all of the **convergence tests** that we use in this class. Every test has certain conditions under which it gives *no answer*, and then you'll have to try a different test. The first few terms are always irrelevant to convergence questions, so every condition only refers to what the terms do *eventually*: at some term a_j and then for every term a_k for $k \geq j$. (I'll write a for the sequence of terms of the series; that is, we are looking at

$$\sum_{n=i}^{\infty} a_n$$

for some integer i, throughout this Section 6.5.)

Every convergence test, if it concludes that a series converges, gives a sequence of approximations of the sum of the series, along with an upper bound on the absolute value of the error of the approximations. Usually, however, we cannot compute the sum of the series exactly. I include here only those estimates (or exact results) that you are expected to know for this course.

The definition

Even the definition of convergence can be viewed as a test. The sequence s in this test always exists; it's the sequence of partial sums in the definition. The problem, however, is that you might not be able to find a nice formula for it!

So, can you find a nice sequence s such that

$$s_m = \sum_{n=i}^{m} a_i$$

(eventually)? If not, then this test gives **no answer**. If so, then go on.

Does

$$\lim_{m\to\infty} s_m$$

exist (as a finite real number)? If not, then the series **diverges**. If so, then the series **converges**.

In fact,

$$\sum_{n=i}^{\infty} a_i = \lim_{m\to\infty} \sum_{n=i}^{m} a_i$$

when this limit converges (by definition).

The Telescoping Series Test

This is a slight variation of the definition that may be easier to spot. Can you find a nice sequence b such that

$$a_n = b_{n+1} - b_n$$

(eventually) or

$$a_n = b_n - b_{n+1}$$

(eventually)? If not, then this test gives **no answer**. If so, then go on.

Does the limit

$$\lim_{n \to \infty} b_n$$

converge (to a finite real number)? If not, then the series **diverges**. If so, then the series **converges**.

In fact,

$$\sum_{n=i}^{\infty} (b_{n+1} - b_n) = \lim_{n \to \infty} b_n - b_i$$

when this limit converges, and

$$\sum_{n=i}^{\infty} (b_n - b_{n+1}) = b_i - \lim_{n \to \infty} b_n$$

when this limit converges.

The Geometric Series Test

Can you write the series as

$$a_n = cr^n$$

(eventually)? If not, then this test gives **no answer**. If so, then go on.

Is $c \neq 0$? If not, then the series **converges**. If so, then go on.

Is $|r| < 1$? If not, then the series **diverges**. If so, then the series **converges**.

In fact,

$$\sum_{n=i}^{\infty} cr^n = \frac{cr^i}{1-r}$$

when $|r| < 1$.

The nth-Term Test

This is probably the first test that you want to consider, unless the series fits one of the special forms above.

Does

$$\lim_{n \to \infty} a_n$$

converge to 0? If not, then the series **diverges**. If so, then this test gives **no answer**.

The Integral Test

Can you find a nice function f defined everywhere (eventually, say defined on $[j, \infty)$) such that $f(n) = a_n$ (eventually)? If not, then this test gives **no answer**. If so, then go on.

Does f take only nonnegative values (eventually)? If not, then this test gives **no answer**. If so, then go on.

Is f monotone decreasing (eventually)? If not, then this test gives **no answer**. If so, then go on.

Does

$$\int_j^{\infty} f(x) \, \mathrm{d}x$$

converge (to a finite real number, for some j)? If not, then the series **diverges**. If so, then the series **converges**.

In this case,

$$\sum_{n=i}^{m} f(n) + \int_{m+1}^{\infty} f(x)\,\mathrm{d}x \leq \sum_{n=i}^{\infty} f(n) \leq \sum_{n=i}^{m} f(n) + \int_{m}^{\infty} f(x)\,\mathrm{d}x,$$

and

$$\sum_{n=i}^{m-1} f(n) + \int_{m}^{\infty} f(x)\,\mathrm{d}x \leq \sum_{n=i}^{\infty} f(n) \leq \sum_{n=i}^{m} f(n) + \int_{m}^{\infty} f(x)\,\mathrm{d}x,$$

for any $m > j$.

The p-Series Test

Can you find a real number p such that

$$a_n = \frac{1}{n^p}$$

(eventually)? If not, then this test gives **no answer**. If so, then go on.

Is $p > 1$? If not, then the series **diverges**. If so, then the series **converges**.

The Direct Comparison Test for Convergence

Does the series consist of only nonnegative terms (eventually)? If not, then this test gives **no answer**. If so, then go on.

Can you find a *convergent* series b such that

$$a_n \leq b_n$$

(eventually)? If not, then this test gives **no answer**. If so, then the original series a also **converges**.

The Direct Comparison Test for Divergence

Can you find a *divergent* series b such that

$$a_n \geq b_n$$

(eventually)? If not, then this test gives **no answer**. If so, then go on.

Does the series b consist of only nonnegative terms (eventually)? If not, then this test gives **no answer**. If so, then the original series a **diverges**.

The Limit Comparison Test

Does the series consist of only nonnegative terms (eventually)? If not, then this test gives **no answer**. If so, then go on.

Can you find a nice series b such that

$$\lim_{n\to\infty} \frac{a_n}{b_n}$$

converges to a *positive* real number? If not, then this test gives **no answer**. If so, then go on.

Does the series b converge? If not, then the original series a also **diverges**. If so, then the original series also **converges**.

The Absolute Convergence Test

Does the series

$$\sum_{n=i}^{\infty} |a_i|$$

of absolute values converge (to a finite real number)? If not, then this test gives **no answer**. If so, then the original series **converges**.

In this case, we say that the original series **converges absolutely**. If the original series converges (which we can only know by some other test) while the series of absolute values diverges, then the original series **converges conditionally**.

The Ratio Test

Does the limit

$$\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|}$$

exist (as a finite real number or infinity)? If not, then this test gives **no answer**. If so, then go on.
Is this limit different from 1? If not, then this test gives **no answer**. If so, then go on.
Is this limit less than 1? If not, then the series **diverges**. If so, then the series **converges**.

The Root Test

Does the limit

$$\lim_{n\to\infty} \sqrt[n]{|a_n|}$$

exist (as a finite real number or infinity)? If not, then this test gives **no answer**. If so, then go on.
Is this limit different from 1? If not, then this test gives **no answer**. If so, then go on.
Is this limit less than 1? If not, then the series **diverges**. If so, then the series **converges**.

The Alternating Series Test

Do we have either

$$a_n = (-1)^n \, |a_n|$$

or

$$a_n = -(-1)^n \, |a_n|$$

(eventually)? If not, then this test gives **no answer**. If so, then go on.
Do we have

$$|a_{n+1}| \leq |a_n|$$

(eventually)? If not, then this test gives **no answer**. If so, then go on.
Does

$$\lim_{n\to\infty} |a_n|$$

converge to 0? If not, then the original series **diverges**. If so, then the original series **converges**.
In this case,

$$\sum_{n=i}^{m} a_n \leq \sum_{n=i}^{\infty} a_n \leq \sum_{n=i}^{m+1} a_n$$

if a_{m+1} is positive, and

$$\sum_{n=i}^{m+1} a_n \leq \sum_{n=i}^{\infty} a_n \leq \sum_{n=i}^{m} a_n$$

if a_{m+1} is negative.

Other tests

There are other tests (and some of these tests can be made more powerful too), but these tests (in these forms) are the only ones that you are responsible for knowing. In particular, every convergence problem in this class should succumb, one way or another, to at least one of these tests. However, there is no end to convergence tests, and mathematicians are still developing new ones, while some series have resisted all efforts so far! (The simplest of these seems to be $\sum_{n=1}^{\infty} n^{-3} \csc^2 n$; you can't tell whether that converges using the tests in this course, but if you can figure it out another way, then you should be able to get a PhD out of it.)

8 Taylor series

One of the major applications of infinite series is to use series to approximate functions that are difficult to calculate. In this class, we mostly concentrate on series that approximate functions that you're already familiar with, because then I can assign you problems that have definite answers. However, the really useful application is when you start with some other problem, such as an integral or a differential equation, that you can't work out exactly using the usual operations but which can still be expressed as an infinite series.

8.1 *Taylor polynomials*

Recall that when a function f is differentiable at a number a, then we can approximate f near a with a linear function that has both the same value and derivative as f does at a:

$$f(x) \approx L(x) = f(a) + f'(a)\,(x-a);$$

here, L is a linear function, $L(a) = f(a)$, and $L'(a) = f'(a)$. This is actually only the beginning (well, slightly after the beginning) of a whole sequence of approximations, each (typically) better than the one before it:

$$\begin{aligned}
f(x) &\approx P_0(x) = f(a); \\
f(x) &\approx P_1(x) = f(a) + f'(a)\,(x-a); \\
f(x) &\approx P_2(x) = f(a) + f'(a)\,(x-a) + \frac{1}{2}\,f''(a)\,(x-a)^2; \\
f(x) &\approx P_3(x) = f(a) + f'(a)\,(x-a) + \frac{1}{2}\,f''(a)\,(x-a)^2 + \frac{1}{6}\,f'''(a)\,(x-a)^3; \\
&\vdots
\end{aligned}$$

(The function that used to be called L is now called P_1.) The general form of this is

$$f(x) \approx P_k(x) = \sum_{n=0}^{k} \frac{1}{n!} f^{(n)}(a)(x-a)^n,$$

where $n! = n(n-1)(n-2)\cdots(3)(2)(1)$. (Recall that $f^{(n)}$ is the nth derivative of f.) Of course, f must be differentiable at a at least k times for P_k to make sense.

The function P_k is the **Taylor polynomial** of f at a of order k. The Taylor polynomial of f at 0 of order k is also called the **Maclaurin polynomial** of f of order k. This terminology is standard (except for some variations in the phrase 'of order' that you may see); however, the notation P_k is *not* standard (and in principle it ought to mention f and a as well as k). Strictly speaking, Taylor polynomials are polynomial *functions* rather than polynomials as such (which are simply algebraic expressions without any variable picked out); otherwise, you'd have to mention the variable x as well.

Notice that a Taylor polynomial P_k of order k really is a polynomial function of degree at most k. (The degree is normally exactly k, but it's smaller if $f^{(k)}(a)$ happens to be 0.) Also, the nth derivative of P_k at a agrees with that of f, if $n \leq k$; that is,

$$P_k^{(n)}(a) = f^{(n)}(a)$$

if $n \leq k$. (On the other hand, if $n > k$, then $P_k^{(n)}(a) = 0$, which is always the case for a higher-order derivative of a polynomial function when the order of the derivative is greater than the degree of the polynomial.) The Taylor polynomial of f at a of order k is the *only* polynomial function of degree at most k whose derivatives at a of order up to k agree with those of f.

Since polynomials are easy to work with, it's convenient to make approximations like these. But in practice, it's also important to know *how good* the approximations are. Since these approximations are based on the behaviour of f at a, we can really only expect them to be good when $x \approx a$. So one way to say that these approximations work is to say that $P_k(x)$ approaches $f(x)$ (or more formally that the error of the approximation, $|P_k(x) - f(x)|$, approaches 0) as x approaches a. This is true for $k = 0$ if f is continuous at a, and for $k > 0$ if f is differentiable k times at a. But in fact, the higher-order Taylor polynomials satisfy a stronger condition:

$$\lim_{x \to a} \frac{|P_k(x) - f(x)|}{|x - a|^k} = 0,$$

which is called (one version of) **Taylor's Theorem**. As x approaches a, $|x - a|$ approaches zero, so dividing by $|x - a|$ would tend to make a positive quantity larger. So P_k is such a good approximation to f that the error not only approaches zero but still approaches zero even after dividing by $|x - a|$ several times.

When investigating these questions, it's helpful to change perspective slightly. Write R_k for $f - P_k$, the Taylor **remainder** of f at a of order k. Then the statement above, showing what a good approximation P_k is, becomes

$$\lim_{x \to a} \frac{|R_k(x)|}{|x - a|^k} = 0,$$

showing how close to zero R_k is. This is good to know, but it may not really be enough; it tells us that moving x close to a will make the approximation better, and very quickly; roughly, when x is already close to a, then moving it twice as close will make the approximation 2^k times better, or you can make the approximation one decimal digit more accurate by moving x only $\sqrt[k]{10}$ times as close. However, this doesn't tell us how accurate the approximation was to start with, nor how close x has to be for this method of improving the approximation to start working.

We can get better results if f is differentiable one more time ($k + 1$ times, not just k times) and near a (not just at a). This strong version of Taylor's Theorem says that

$$R_k(x) = \frac{(x - a)^{k+1}}{k!} \int_{t=0}^{1} (1 - t)^k f^{(k+1)}(a - at + xt) \, \mathrm{d}t,$$

as long as f is continuously differentiable $k + 1$ times (at least) between a and x. (The integral here may exist even if f is not *continuously* differentiable $k + 1$ times, but then the value of this integral might not equal the remainder.) To be more explicit, here is the statement for the first few values of k:

$$\begin{aligned} f(x) &= f(a) + (x - a) \int_{t=0}^{1} f'(a - at + xt) \, \mathrm{d}t \\ &= f(a) + f'(a)(x - a) + (x - a)^2 \int_{t=0}^{1} (1 - t) f''(a - at + xt) \, \mathrm{d}t \\ &= f(a) + f'(a)(x - a) + \frac{1}{2} f''(a)(x - a)^2 + \frac{(x - a)^3}{2} \int_{t=0}^{1} (1 - t)^2 f'''(a - at + xt) \, \mathrm{d}t \\ &\;\;\vdots \end{aligned}$$

These statements may be proved by repeated application of integration by parts (and the Fundamental Theorem of Calculus, which is why $f^{(k+1)}$ must not only exist but also be continuous). To be specific, you can prove each statement using $u = (1 - t)^k \big/ k!$ and $v = (x - a)^k f^{(k)}(a - at + xt)$, integrating by parts, simplifying, and (if applicable) applying the previous statement.

For purposes of approximation, it's useless to actually work out the integral that appears here; if you knew the exact value of $f^{(k+1)}$ at all of the points between a and x, then you could probably just evaluate f at x directly. However, if there is a value M_k such that you know that $f^{(k+1)}$ never has an absolute value greater than M_k at any point between a and x, then you can use M_k to get a bound on the remainder:

$$|R_k(x)| \le \frac{M_k}{(k + 1)!} |x - a|^{k+1}.$$

The reason for this is that we know that $R_k(x)$ is exactly the integral that appeared in the full version of the theorem, and we can bound its absolute value using the bound on its integrand:

$$\begin{aligned}|R_k(x)| &= \left|\frac{(x-a)^{k+1}}{k!}\int_{t=0}^{1}(1-t)^k f^{(k+1)}(a-at+xt)\,\mathrm{d}t\right| \\ &\leq \frac{|x-a|^{k+1}}{k!}\int_{t=0}^{1}(1-t)^k \left|f^{(k+1)}(a-at+xt)\right|\mathrm{d}t \\ &\leq \frac{|x-a|^{k+1}}{k!}\int_{t=0}^{1}(1-t)^k M_k\,\mathrm{d}t = \frac{|x-a|^{k+1}}{k!}\frac{M_k}{k+1} = \frac{M_k}{(k+1)!}|x-a|^{k+1}.\end{aligned}$$

To be more specific:

$$|R_0(x)| = |f(x) - f(a)| \leq M_0\,|x-a|$$

if $|f'|$ is never greater than M_0 between a and x,

$$|R_1(x)| = \left|f(x) - \big(f(a) + f'(a)(x-a)\big)\right| \leq \frac{1}{2}M_1\,|x-a|^2$$

if $|f''|$ is never greater than M_1 between a and x,

$$|R_2(x)| = \left|f(x) - \left(f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2\right)\right| \leq \frac{1}{6}M_2\,|x-a|^3$$

if $|f'''|$ is never greater than M_2 between a and x, etc. Note that this upper bound on the absolute value of the remainder is basically the absolute value of the next term that you would add if you went one step further, except that instead of using a derivative at a, you must use the largest derivative (in absolute value) anywhere between a and x.

8.2 *Power series*

We can extend from polynomials to infinite **power series** and get the **Taylor series** of f at a:

$$P_\infty(x) = \sum_{n=0}^{\infty}\frac{1}{n!}f^{(n)}(a)(x-a)^n.$$

(When $a = 0$, this is the **Maclaurin series** of f.) This power series exists as long as f is infinitely differentiable at a, that is as long as f has derivatives of all orders at a. However, there are no theorems guaranteeing that this series converges, nor that it's anything like $f(x)$ when it does converge (except that it must converge to $f(a)$ when $x = a$ exactly). We say that f is **analytic** at a if this series converges to $f(x)$ at least on some interval around a. Any function built out of the usual operations* is analytic, as long as it's infinitely differentiable, so everywhere that it is defined except where an absolute value or a root (or a power with a fractional exponent) is applied to 0 or an inverse trigonometric sine, cosine, secant, or cosecant is applied to ±1. However, there are functions for which the Taylor series exists but fails to converge (except when $x = a$ exactly); the only examples that I know are defined themselves as series, such as $f(x) = \sum_{n=0}^{\infty}\mathrm{e}^{-\sqrt{2^n}}\cos(2^n x)$ (which is not a power series but still converges everywhere by the Root Test). There are also functions for which the Taylor series converges but not to $f(x)$ (except when $x = a$ exactly); an example of this (with $a = 0$) is $f(x) = \begin{cases}\mathrm{e}^{-x^2} & \text{for } x \neq 0, \\ 0 & \text{for } x = 0.\end{cases}$ But as you can see, neither of these

* addition, subtraction, multiplication, division, taking opposites, taking reciprocals, taking absolute values, raising to the power of a constant, raising to a power when the base is positive, taking roots with a constant index, taking roots with a positive radicand, taking logarithms, the six trigonometric operations, and the six inverse trigonometric operations

examples is built entirely out of the usual operations; the first also involves a limit (to sum the infinite series), while the second is defined piecewise.

There are several famous Taylor series of analytic functions that you should know:

$$\begin{aligned}
x^k &= \sum_{n=0}^{\infty} \binom{k}{n} (x-1)^n \text{ for } 0 < x < 2; \\
e^x &= \sum_{n=0}^{\infty} \frac{x^n}{n!}; \\
\ln x &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n+1} (x-1)^{n+1} \text{ for } 0 < x \leq 2; \\
\sin x &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}; \\
\cos x &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}; \\
\operatorname{atan} x &= \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1} \text{ for } -1 \leq x \leq 1.
\end{aligned}$$

You can check that these are Taylor series for the claimed functions by checking the functions' derivatives, and you can prove that these series converge for the claimed values of x using the usual convergence tests, but it takes more work to prove that they converge *to* the claimed functions. (Much of this is proved in the textbook in Sections 9.7–9.10, or you can trust the theorem that the usual operations are all analytic wherever they're differentiable.)

The formula for x^k may seem particularly useless, and it mostly is when k is a whole number, but it is valid for any real number k, such as $k = -1$ (for $1/x$), $k = 1/2$ (for $\sqrt{x}$), etc. This formula includes $\binom{k}{n}$, the **binomial coefficient** of k with index n, which is defined by

$$\binom{k}{n} = \frac{k^{\underline{n}}}{n!} = \frac{k(k-1)(k-2)\cdots\big(k-(n-1)\big)}{n(n-1)(n-2)\cdots 1};$$

that is, the binomial coefficient is a fraction whose numerator and denominator each consists of n factors, with the denominator beginning at n to produce $n!$ and with the numerator beginning at k to produce $k^{\underline{n}}$, the **falling power** of k with index n (so in particular, $n! = n^{\underline{n}}$). Just as $0! = 1$, so $\binom{k}{0} = \frac{1}{1} = 1$; another useful fact is that $\binom{-1}{n} = (-1)^n$. (There is really a lot to be said about this stuff, which is part of the branch of mathematics called *combinatorics*, but the only thing that you're responsible for is to calculate $n!$ and $\binom{k}{n}$ for specific values of k and n.)

When you use these formulas, you may need to substitute some other expression for x, and you may need to start a sum at some other index. For example, if you want to evaluate

$$\sum_{n=3}^{\infty} \frac{x^n}{n},$$

then the important thing to notice is that the denominator is the same as the exponent (rather than the factorial of the exponent, as in some of the formulas) and that almost every natural number appears as an exponent (rather than only odd numbers or only even numbers, as in some of the formulas), which means that it's the formula for $\ln x$ that's relevant. To get the exponent in the right form, choose m so that $n = m + 1$; that is, $m = n - 1$. You now have

$$\sum_{m=2}^{\infty} \frac{x^{m+1}}{m+1}.$$

To get the right base, you might choose y so that $x = y - 1$; however, to get the factor of $(-1)^n$ as well, you should actually choose y so that $x = -(y-1)$. That is, $y = 1 - x$, so you now have

$$\sum_{m=2}^{\infty} \frac{\big(-(y-1)\big)^{m+1}}{m+1} = \sum_{m=2}^{\infty} \frac{(-1)^{m+1}}{m+1}(y-1)^{m+1} = -\sum_{m=2}^{\infty} \frac{(-1)^m}{m+1}(y-1)^{m+1}.$$

Now you can match this against the formula for $\ln x$, using m in place of n and y in place of x, with an extra minus sign out front and with the first two terms missing. Since these missing terms are

$$\sum_{m=0}^{1} \frac{(-1)^m}{m+1}(y-1)^{m+1} = \frac{1}{1}(y-1)^1 + \frac{-1}{2}(y-1)^2 = -\frac{1}{2}y^2 + 2y - \frac{3}{2},$$

the original series equals $-\big(\ln y - (-1/2\,y^2 + 2y - 3/2)\big) = -\ln y - 1/2\,y^2 + 2y - 3/2$ whenever $0 < y \leq 2$. Remembering that $y = 1 - x$, you can finally conclude that

$$\sum_{n=3}^{\infty} \frac{x^n}{n} = -\ln(1-x) - \frac{1}{2}(1-x)^2 + 2(1-x) - \frac{3}{2} = -\ln(1-x) - \frac{1}{2}x^2 - x \text{ for } -1 \leq x < 1.$$

Some of these formulas appear in slightly different forms in the textbook; one version may be more convenient for a particular problem than another, but either version should suffice for all of the relevant problems.